Psychological Bulletin

HARRY HELSON, Editor
Kansas State University

CONTENTS

Published Bimonthly by the
American Psychological Association

VOL. 59, NO. 4 July 1962

The *Psychological Bulletin* contains evaluative reviews of research literature and reviews of research methodology and instrumentation in psychology. This JOURNAL *does not publish reports of original research or original theoretical articles.*

Abstracts. Beginning with the January 1963 issue, all articles will be preceded by an abstract of 100 to 120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts.* Detailed instructions for preparation of the abstracts may be obtained from the Editor or from the APA Central Office.

Manuscripts should be sent to the Editor, Harry Helson, Department of Psychology, Kansas State University, Manhattan, Kansas.

Preparation of articles for publication. Authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (1957 Revision). Special attention should be given to the section on the preparation of the references (pp. 50–60), since this is a particular source of difficulty in long reviews of research literature. *All copy must be double spaced, including the references.* All manuscripts should be submitted *in duplicate,* one of which should be an original typed copy; author's name should appear only on title page. Dittoed and mimeographed copies are not acceptable and will not be considered. Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Reprints. Fifty free reprints are given to contributors of articles and notes.

Communications—including subscriptions, orders of back issues, and changes of address—should be addressed to the American Psychological Association, 1333 Sixteenth Street N.W., Washington 6, D.C. Address changes must reach the Subscription Office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Annual subscription: $10.00 (Foreign $10.50). Single copies, $2.00.

Published Bimonthly by:
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
Menasha, Wisconsin
and 1333 Sixteenth Street N.W., Washington 6, D.C.

Second class postage paid at Washington, D.C., and at additional mailing offices. Printed in U.S.A.

VOL. 59, NO. 4 JULY 1962

Psychological Bulletin

THEORIES OF VIGILANCE[1]

JUDITH P. FRANKMANN AND JACK A. ADAMS
University of Illinois

Vigilance research concerns the attentiveness of the subject and his capability for detecting changes in stimulus events over relatively long periods of sustained observation. Interest in this topic has accelerated rapidly and the volume of experimental findings has increased steadily in recent years. With investigators spread over several continents and publishing under the sponsorship of numerous military, industrial, and academic organizations, it has become a major problem to keep up with the technical literature. This report is a critical survey of the existing literature, with emphasis being given to the organization of experimental results under the several theoretical hypotheses which have been advanced in explanation of the findings. The many diverse sources of technical papers made complete coverage of the literature difficult, but it is believed that only a small fraction of the papers relevant to contemporary theories were unavailable.

The increased number and productivity of researchers has been associated with a greater variety of experimental situations. Those covered in this paper generally are one of four types: (*a*) Classical vigilance tasks, e.g., the Mackworth Clock Test (Mackworth, 1950). Near-threshold transient critical signals are randomly presented against a background of neutral signals. (*b*) Multiple display situations where a critical signal could occur at any one of several stimulus sources, e.g., Broadbent's Twenty Dials Test (Broadbent, 1950). Constant scanning of the several stimulus sources is required. (*c*) Threshold measurement, e.g., Bakan (1955). A train of signals is presented, starting at random intervals in time, with an intensity increment at each step until the observer detects the signal. (*d*) Observing response experiments, e.g., Holland (1957, 1958), where visual attending is measured indirectly through some other response that suggests observing of the stimulus display. Frequency of observing is then related to percentage detection of the critical signal occurring on the display.

Although many experiments were generated by specific practical questions, a framework for organizing the accumulation of empirical findings has not been neglected. It is now possible to distinguish a number of explanatory systems. The main purpose of this paper is to review these

[1] This paper was prepared under Contract AF 19(604)-5705 between the Aviation Psychology Laboratory, Department of Psychology, University of Illinois, and the Operational Applications Laboratory, Deputy for Technology, Electronic Systems Division, Air Force Systems Command. A version of this paper was published by the monitoring agency under the same title as Technical Note AFCCDD-TN-60-25.

models and discuss them in terms of their effectiveness in accounting for the empirical findings.

Inhibition

Mackworth (1950) advanced the first comprehensive interpretation of vigilance behavior relating observed phenomena of watch-keeping to principles of Pavlovian classical conditioning. In the same report Mackworth presented extensive data on deterioration of criterion performance on a number of different tasks under conditions of prolonged monitoring—the Clock Test, the Synthetic Radar Test, and the Auditory Listening Test. The Clock Test, used most extensively, had a blank circular face with a hand that moved one step each second. Occasionally the hand moved in a double step, according to a prearranged schedule, and this was the critical signal to be detected and reported by pressing a response key. Over a 2-hour observation period it was found that the percentage of critical signals detected was a decreasing negatively accelerated function of time, with the greatest drop occurring during the first half-hour. The analogy drawn with classical conditioning was that original conditioning took place in the demonstration period where the conditioned stimulus was the double jump of the clock hand and the unconditioned stimulus was the experimenter's informing comment "Now!" The conditioned voluntary response was the subject's pressing the key to the double jump. Knowledge of results is a reinforcing state of affairs. The 2-hour observation period then, was considered an extinction period where the unconditioned stimulus and reinforcement provided by the experimenter was absent. During extinction the percentage of detections declined, and Mackworth attributed this decline to the growth of internal inhibition. Other evidence for interpreting vigilance data in terms of Pavlovian conditioning was the temporary but complete restoration of initial performance level by the occurrence of a telephone message to the observer in the middle of the watch-keeping session. Mackworth viewed this as an instance of disinhibition where an alien stimulus produced a temporary increase in responsiveness. Other evidence for a classical conditioning interpretation occurred in an experiment when the experimenter provided knowledge of results after each double jump of the clock hand and prevented the occurrence of a decrement in detection. This, within Mackworth's explanatory frame of reference, would be a reinforcing operation and would be expected to keep the performance level high.

Mackworth was obliged to qualify a strict interpretation in terms of classical conditioning because he was unable to obtain anything near complete failure of responding, i.e., total experimental extinction. In fact, for the Clock Test, the level of detection for the critical signal ordinarily stabilized at about 70–75%. A state of expectancy and self-instructions were hypothesized as partly replacing the unconditioned stimulus and its reinforcing function.

The inhibition analysis of vigilance behavior came with a ready-made development so that the theorist's task has been one of coordinating aspects of vigilance behavior with conditioning phenomena. From the standpoint of handling more recent results, inhibition does not fare very well. For example, a high frequency of signals should result in a greater vigilance decrement than low frequency signals because, within the classical conditioning framework of

the inhibition hypothesis, it represents a relatively high frequency of extinction trials. Yet, Deese and Ormond (1953) and Jenkins (1958) show the opposite to be true. Mackworth (1957) himself has come to regard the expectancy explanation as more important in accounting for recent findings, but he was not ready to dispense with reinforcement and nonreinforcement effects completely. Although the inhibition hypothesis plausibly accounted for many of the observed results of Mackworth's experiments, it never gained wide acceptance. Mackworth's research rather than his interpretation has been responsible for generating new experiments. Reluctance to accept the inhibition explanation also has been based on attitudes towards theory construction. For example, Deese (1955, p. 366) felt it unnecessary to postulate separate inhibitory and excitatory processes when a single state of vigilance which declines under specified conditions will handle the data as well.

Attention

Broadbent (1953) carried the analogy of watch-keeping to Pavlovian conditioning even further than Mackworth. However, instead of interpreting vigilance in terms of conditioning he interpreted both in terms of attention. He contends that the organism will select stimulus subsets from the impinging stimuli because (*a*) the nervous system cannot handle the total volume of stimulation at any given instant, and (*b*) adequate responding to one part of the stimulus situation is incompatible with adequate responding to another part. At least three properties of stimuli are important in determining priority of selection. *Physically intense* stimuli are more apt to be selected than weak stimuli. Stimuli of greater *biological importance* at the moment have higher priority of selection. Finally, *novel stimuli*, i.e., those differing more from immediately preceding stimuli, have an increased likelihood of being selected.

The decrement in accuracy of detection is accounted for by the competition of stimuli in Broadbent's view. The repeated application of a stimulus results in reduced novelty, allowing other parts of the stimulating situation to gain priority. Decrements over time have been reported for a variety of response measures. The following references are representative, not exhaustive: probability of detection (Adams, 1956; Bakan, 1952, 1953, 1955, 1957; Deese & Ormond, 1953; Jenkins, 1958; Jerison, 1958, 1959; Jerison & Wallis, 1957a, 1957b; Kappauf & Powe, 1959; Mackworth, 1948, 1950, 1957), response time (Garvey, Taylor, & Newlin, 1959; McCormack, 1958), threshold intensity (Bakan, 1955; Garvey, Henson, & Gulledge, 1958; McFarland, Holway, & Hurvich, 1942).

During a rest period different stimuli are selected allowing the original task stimuli to regain novelty. This corresponds to the observed improvement in detection following rest. Mackworth (1950) and Jenkins (1958) found an increase in probability of detection with interspersed rest periods. Adams (1956) reported recovery of decrement following a 10-minute rest. Similar recovery was found in response time (McCormack, 1958) and luminance threshold (McFarland et al., 1942) following rest.

In similar fashion, a new stimulus introduced between applications of the original stimulus will temporarily renew the novelty of the original one since it is then different from the immediately preceding stimulus. Mackworth's telephone message would fit

this category. McFarland et al. (1942) observed that forced conversation with an observer after 105 minutes of measuring luminance threshold produced a marked but temporary increase in sensitivity. Body stretching had a similar effect and can be interpreted as new stimulation. The same reasoning applies when knowledge of results is given following each critical signal; the novelty of the signal is maintained (Baker, 1959b; Mackworth, 1950; Pollack & Knaff, 1958). We see then that in Broadbent's model, disinhibition and reinforcement are examples of the same phenomenon.

When several sources must be monitored some of them have higher priority initially, but as the watch period progresses attention shifts towards previously neglected sources. The overall level of performance is maintained, but becomes irregular. In an experiment using 20 dials as signal sources Broadbent (1950) found this to be the case. Loeb and Jeantheau (1958) reported no decrement in a 20 dials test; Howland (1958) found the same result using four meters. With a three-clock display, detection level remained stable (Jerison & Wallis, 1957a; Jerison & Wing, 1957) but in comparison with a one-clock display Jerison and Wallis (1957a) found overall detection level was much lower. A fine-grain analysis in the same report hinted that a decrement may have occurred in just the first 3–4 minutes of the watch with three clocks, but this is quite a different order of phenomenon than the large decrement for a one-clock test unit that develops over a relatively long time period.

The critical signal itself serves as a novel stimulus and partially restores performance. Broadbent used this to explain Mackworth's (1950) finding that observers who detected more signals maintained a higher level of vigilance throughout the watch. This interpretation was also related to the fact of better performance with higher signal rates (Deese & Ormond, 1953; Garvey et al., 1959; Jenkins, 1958; Kappauf & Powe, 1959).

A continuous intense irrelevant stimulus will not initially have much effect. As the watch continues and the original stimulus loses novelty any decrement will be accentuated. This is more likely with a single stimulus source. Broadbent (1954) had observers monitor 20 dials over a 5-day period with noise on Day 3 and Day 4. Under noise conditions there were significantly fewer responses made in 9 seconds or less, than in quiet conditions. Using 20 lights he found no significant difference, but the lights were more noticeable signals. Loeb and Jeantheau (1958) also using 20 dials reported longer response latencies throughout the watch with noise and vibration, but no changes with time. In a three-clock test, Jerison and Wing (1957) introduced noise for $1\frac{1}{2}$ hours following $\frac{1}{2}$ hour of quiet. A decrement developed in the final half-hour. With only one clock, Jerison and Wallis (1957b) found no effect of noise relative to quiet conditions.

Broadbent's model consists essentially of the assumption that selection of stimuli is necessary in accordance with the three stimulus properties given above. The formal development of the model was not carried beyond these broad assumptions. Although Broadbent did not intend to cover all of the facts of vigilance, even the results he did consider seem more specific than the model can convincingly handle. Application to the general trends such as decrement with prolonged watch, recovery of performance with rest, and maintained efficiency with continued knowledge

of results appeared grossly to follow from the principles of stimulus selection. In other cases where the experimental situation was subjected to finer analysis it was difficult to see how Broadbent's interpretations followed from the model. A number of his more detailed applications seemed like ad hoc explanations of known results rather than predictions deduced from the original principles.

Expectancy

The expectancy hypothesis of vigilance was originally proposed by Deese (1955). He began with the notion of an excitatory state of vigilance which determines the probability of detection for any observer. The expectancy hypothesis states that:

> (*a*) the observer's expectancy or prediction about the search task is determined by the actual course of stimulus events during his previous experience with the task, and (*b*) the observer's level of expectancy determines his vigilance level and hence his probability of detection (p. 362).

It should be emphasized that the second part of the hypothesis does not, for Deese (1955), imply that level of vigilance is directly determined by expectancy (pp. 364–365). The level of vigilance for any observer is also subject to modification due to changes in his motivational states whereas his extrapolation of future stimulus events might not be affected by such changes. Deese wanted to avoid the artificial situation where expectancy completely determined vigilance. These states, according to Deese, are the basis of individual differences in vigilance and it is the psychologist's task to discover measures of behavior which predict levels of vigilance expected of an individual in a search task. But do these nonexpectancy states serve merely to raise or lower the probability of detection by some constant amount throughout the task, or is the form of the detection curve over time changed as well? Deese does not clarify this matter too well, but a free interpretation of his exposition on vigilance (Deese, 1955) would suggest that the nonexpectancy states determine a base level for an individual's vigilance. Expectancy, however, determines both the overall level and the short range variations in probability of detection. It is assumed that the average level of expectancy, and thus detectability, is a positive function of signal rate, while the short range variations in expectancy are determined by the ongoing intersignal interval. Deese assumes that expectancy is determined by all of the past stimulus events in the task and he elaborates this notion by relating expectancy to intersignal interval and stating that it increases up to the value of the mean intersignal interval and beyond. Thus, it would appear that probability of detection would be below average when an intersignal interval is less than the mean interval, and equal to or greater than the average probability of detection when the intersignal interval is equal to or greater than the mean.

About the only evidence that can be found for the expectancy hypothesis is that probability of detection is a positive function of signal rate (Deese & Ormond, 1953; Jenkins, 1958). However, little or no evidence can be found in support of Deese's views of expectancy as a function of intersignal interval. In analysis of some of his own data (Deese & Ormond, 1953), Deese found little effect of interval size, although a slight tendency for higher probability of detection for longer intervals could be considered small support for the expectancy prediction. Analysis of intersignal interval has not led to any consistent results even yet. Jerison

and Wallis (1957a) and McCormack (1958) found no effect of interval size. Jenkins (1958) reported detection dropped monotonically with increasing intervals when average rate was as high as 480 per hour; at lower rates he found no effect. Bartlett, Beinert, and Graham (1955) found lower probability of detection with longer intervals using a 40 per hour signal rate. Mackworth's data showed better detection for brief intervals than for his longer 10-minute intervals. Kappauf and Powe (1959) reported a U shaped function in an audio-visual checking task. One can scarcely imagine a more varied set of results and considering average rates and ranges of time intervals does not resolve the conflict among these data. Jenkins (1958) suggested that the average rate of signals has a much greater effect on detection level than short range fluctuations, so the issue becomes less critical from a practical standpoint. Harabedian, McGrath, and Buckner (1960) emphasize that for a basic understanding, a major methodological problem exists in defining an intersignal interval because it can be expressed in terms of (*a*) time between signals, whether the signal is detected or not; (*b*) time since the last detected signal; and (*c*) time since the last missed signal. Their results from audio and visual vigilance tasks revealed differences dependent upon the method chosen to define the interval.

Baker (1958, 1959a, 1959b, 1959c) has elaborated Deese's expectancy hypothesis and has provided a body of experimental evidence in support of his own views. A major portion of Baker's arguments in applying the expectancy model to experimental variables rests on the single consideration that an operator's expectancy is based on how he perceives the actual series of stimulus events. Any variation which makes confirmation of expectancy more likely or which allows more accurate perception of the actual stimulus events should lead to better performance. For example, when a signal is missed and the observer is unaware of the omission, the intersignal interval is increased and expectancy is lowered. The concern of Harabedian et al. (1960) with problems of defining intersignal interval would seem to be close to Baker's interests here.

Operationally, Baker's expanded definition of expectancy has five major classes of variables:

Average signal rate. Baker's primary interests have been in predicting short range variations in detection as a function of intersignal interval, but he would seem to agree with Deese and Jenkins that detection probability is a positive function of average signal rate (Baker, 1959c).

Regularity and range of the intersignal interval. Regularity of the signal increases the probability that the expectancy state will be reinforced. Baker contends that expectancy grows as the interval following a signal increases to the value of the mean intersignal interval and, beyond the mean value, expectancy falls to a low level. Notice that this is a modification of Deese's view (1955). Baker (1959b) tested this hypothesis in a reaction time experiment similar to that of Mowrer (1940). He measured button pressing response times to an initial series of 20 light signals and then varied the interval before the final twenty-first signal. Using a 2×2 factorial design the initial series was presented at 10-second or 2-minute mean intervals with regular or irregular intersignal intervals. Following the regular 10-second series, changes in reaction time to the twenty-first signal paralleled the predicted course. For very short inter-

vals reaction time was long and following a decrease as the mean interval was approached reaction time again became somewhat longer up to the highest value tested (30 seconds), but not nearly as long as the reaction time for short intervals. The irregular and longer interval condition showed little variation in reaction time to the twenty-first signal although when a trend appeared it tended to support the expectancy hypothesis. This experiment suggests that in vigilance tasks, where the signals are always irregular and usually occur at low average rates, one should not expect to find a large effect of intersignal interval.

The range of intersignal intervals was related to the occurrence of decrement, and apparently Baker has been the first to demonstrate this phenomenon. Using a simulated PPI display, Baker (1958) found no decrement when the intersignal intervals ranged from 36–196 seconds, but in a later study (Baker, 1959a), using the same task, found a decrement when the interval range was increased to 45–645 seconds. Interestingly, this latter range of intervals was that used by Mackworth in generating his well-known decrements in detection with several different visual and auditory monitoring tasks. In another experiment, using a simulated B-scan radar display, Baker (1959b) assessed the effects of complete signal regularity with an occurrence every 2½ minutes, a random series with a range of intervals from 1 to 6 minutes, and a wide range of intervals where the spread was ¾–10 minutes (randomly arranged). Signal frequency was the same for all groups, being 24 an hour. Decrement was found only for the group with the widest range of intersignal intervals. Baker(1959c) would interpret this as the subject abandoning any efforts to form expectancies because it is done too imprecisely when intervals are long.

Knowledge of results. This variable prevents a decrement by allowing an accurate perception of the sequence of stimuli. Baker (1959b) tested three groups of subjects given (*a*) no information; (*b*) complete information on correct, missed, and false signals; and (*c*) repetition of a missed signal at five-second intervals until detected. The task was visual detection, the observation time 1 hour, and the signal rate 24 an hour. Only the group with no feedback had a significant decrement. Mackworth (1950) earlier had found that informing an observer of his success or failure in detecting a signal served to completely eliminate the decrement in detection. Pollack and Knaff (1958) obtained results similar to Mackworth's.

Knowledge of signal location on a visual display. Knowledge of signal location makes confirmation of expectancy more likely. With increasing variability in location the appropriate part of the search area may not be scanned when the signal occurs. This leads to a lower apparent signal frequency and lower probability of detection. While Baker is concerned only with spatial variables as they influence temporal expectancy, Mackworth (1950, pp. 58–59) implies a spatial expectancy for signal occurrence at one or more locations as a state distinct from temporal expectancy. Deese and Ormond (1953) varied the distribution of signals on a radar display presenting 50% in one quadrant. Detection in the high probability quadrant was only slightly superior to the other three during an hour period, but the overall probability was very high. In a similar experiment, Nicely and Miller (1957) found greater detection in the more frequent quadrant, the difference

increasing in the last half-hour of the 74-minute watch. Level of detection in the high probability area remained relatively constant throughout. Bartlett et al. (1955) used the method of constant stimuli to measure brightness thresholds. Knowing the time but not the location of a signal led to higher thresholds than knowing both the time and place of occurrence. A much larger decrement in performance was observed when neither time nor location was known. Krendel and Wodinsky (1959) measured time to detect a randomly located light signal when the time of onset was known. They found no decrement in search time over an hour. Finally, Garvey et al. (1958) reported that no increase in stimulus intensity was necessary to detect a signal appearing after 60 minutes of monitoring provided the observer was warned of signal time and location before it appeared. Without such knowledge observers showed a large increase in visual threshold.

Signal intensity. Baker holds that expectancy is more likely to be confirmed with more intense signals. Both Adams (1956) and Mackworth (1950) found higher probability of detection for visual signals of higher intensity than lower intensity. And, if we assume that perceived signal intensity is related to the duration of the signal, Adams (1956) found a higher probability of detection for signals of 2 seconds in length than for signals of 1 second.

Looking at the combined efforts of Deese and Baker in forwarding the expectancy hypothesis we again have a theory at the early stages of development making qualitative predictions about vigilance behavior. The underlying assumptions were set forth more clearly than was the case with Broadbent's attention hypothesis with the result that the expectancy hypothesis lends itself more readily to testing. A case in point is Baker's report (1959b) initiated with the intention of evaluating the model. The expectancy hypotheses do not grapple explicitly with the classical vigilance issue of decrement accruing over observation time, which might be listed under *long range effects*, and distinguished from *short range effects* where momentary determiners of response (intersignal interval, spatial location of the signal, etc.) are emphasized. These latter effects intrigue expectancy theorists. Other variables known to be important are largely neglected by Deese and Baker, e.g., rest periods and environmental factors such as presence of the experimenter, interpolated messages, and noise. Oddly enough, Deese (1955) devoted some space to the importance of varied background sensory input in maintaining vigilant behavior but he did not relate it to expectancy and thus might be said to have a two-factor theory. Baker (1959c) mentioned environmental factors as possible distractions that would lower apparent signal frequency, but these factors were not formally entered into his theory.

Varied Sensory Environment

Scott (1957) explored Hebb's thesis (1955) that stimuli serve a dual function: (*a*) they have a cue function in controlling goal responses (the function usually ascribed them in learning theories), and (*b*) an arousal or vigilance role to which Hebb (1955) ascribes motivational properties. Scott feels that the arousal function of stimuli has been largely ignored and should be given more attention. Broadbent (1958) calls this the "activationist hypothesis" in vigilance research, and his views (Broadbent, 1953) on stimulus variety were somewhat similar. To document

implications of arousal for vigilance, Scott surveyed the literature concerned with performance deterioration in a variety of repetitive tasks with particular attention to the uniformity of sensory environment that accompanied such activities. He concluded that loss of efficiency was directly related to reduction in stimulus variation. When background stimuli are at a minimum and only occasional and often low key critical stimuli are present, rapid deterioration should be expected. The more unchanging are the critical stimuli, the sooner deterioration will occur. Rest periods and introduction of extraneous stimuli serve to increase the variety of stimulation needed to maintain or restore efficient behavior.

Neurophysiological as well as behavioral research support the importance of a secondary role for stimuli. Impulses from the same sensory stimuli have been shown to reach the cerebral cortex via two different pathways. They travel directly along the sensory tract to the corresponding nucleus in the thalamus and terminate in a specific projection area of the cortex. A second pathway has been studied, wherein impulses from the same stimuli travel a slow circuitous route through the ascending reticular activating system which discharges a diffuse bombardment over wide areas of the cerebral cortex. The latter type of cortical stimulation is considered necessary for the maintenance of alert behavior. Scott (1957), Hebb (1955), Lindsley (1957), Malmo (1959), and Samuels (1959), summarize the experiments related to this work.

Given the nonspecific effect of stimuli on behavioral organization, Scott suggested that stimuli lose their nonspecific effects with continued exposure, the rate of such habituation increasing as the environment is more uniform. This process, termed "sensory habituation," results in a wide range of modifications in behavior of which loss of vigilance is one of the earliest to appear. Under conditions of severe isolation over extended periods more serious symptoms such as hallucinations appear, as in the McGill studies of sensory deprivation.

The sensory habituation theory finds application to vigilance tasks in a number of ways. One would expect to find performance restored to or maintained at a higher level under conditions which increase the variety of either peripheral or relevant task stimuli. Examples cited by Scott (1957) included: rest periods, high signal rate, knowledge of results, interpolated messages, use of tasks with multiple stimulus sources, and presence of the experimenter. Data relevant to most of these factors has been summarized in earlier sections, with the work of McFarland et al. (1942) being particularly relevant. The results from vigilance tasks with multiple stimulus sources support Scott's position quite well and, in fact, his is the only successful theory in this vein (Broadbent, 1950; Hoffman & Mead, 1943; Howland, 1958; Jerison & Wallis, 1957a, 1957b; Jerison & Wing, 1957; Loeb & Jeantheau, 1958). All of these studies have the distinctive feature of showing no vigilance decrement whatsoever—a puzzling but consistent finding that has received little systematic attention. Most investigators have chosen to use tasks where decrement is known to occur and largely have ignored the potential for understanding vigilance behavior that might be found in studying the tasks that fail to yield detection decrement. For example, one might entertain the hypothesis that the sensory inputs

sustaining responsiveness might arise from the proprioceptive stimulation derived from head and eye movements. It appears that under some experimental conditions, not yet clearly defined, task complexity and variety eliminates vigilance decrements in most cases. A negative instance is Garvey et al. (1959), where a decrement was found with a multidial task using a very low signal rate. Their task was somewhat different from the conventional vigilance task because they required the observer to detect a larger deviation occurring to a constantly moving needle in each dial. The very low signal rate (average of 2.5/2 hours) may be the reason for the difference because Howland (1958) used the same general type of task and found no decrement.

Scott was not proposing a theory of vigilance in his paper, but the relevance of his view to this area is clear. He provided convincing evidence for the presence of perceptual variation as a necessary condition in maintaining alertness. Although Mackworth (1950) and Deese (1955) have noted the importance of such variation, this point has not been formally incorporated in any of the models dealing specifically with vigilance. It is worthy of attention.

Observing Responses

The analysis of vigilance behavior in terms of rate of observing responses is not a theory but a technique for studying vigilance. Theory enters the picture only in the assumption that detection of a signal serves as a reinforcement for the observing response (Holland, 1958). Holland (1957, 1958) has been the major promoter of this type of analysis. His purpose was to show the influence of schedules of reinforcement on rate of observing response and the parallel influence on detection performance. The extent to which rate of observing and probability of detection follow the same course would determine how much of the detailed knowledge about schedules of reinforcement (e.g., Ferster & Skinner, 1957) could be carried over directly to vigilance behavior. The observing response studied by Holland was that of pressing a key to illuminate a dial. The pointer on the dial deflected from the null position at intervals set by the schedule of reinforcement and remained deflected until the observer reset the pointer by pressing a second key. The observer could only see the dial by pressing the key.

Holland studied rate of observing as a function of several common reinforcement schedules to test the assumption that detection serves as a reinforcement. On fixed interval schedules ranging from ½–4 minutes, observers learned temporal discriminations reflected by "scallops" in the cumulative response curves during the last of eight 40-minute sessions. During extinction the rate remained high for a time and then gradually decreased to a low level.

Following a fixed ratio schedule with ratios increasing from 36–200 responses per reinforcement, Holland reported a higher rate of observing with higher ratios. Extinction curves were typical in showing spurts of high responding in a jagged decline in rate. These results along with successful training on multiple schedules and responding at low rates led Holland to conclude that signal detection could serve as a reinforcement for observing responses. His next step was to use schedules of signal presentation identical to those in typical vigilance tasks.

Rate of observing was measured on variable-interval schedules with average intervals of 15 seconds, 30 seconds, 1, 2, and 3 minutes. These covered

the range in terms of signal rate from 20–240 signals per hour. Observing rate was higher with higher signal rates. The 3-minute interval led to a decline in rate over time paralleling the decrement in percentage of signals detected reported by Deese and Ormond with a corresponding 20 signals per hour. Among the other values studied, both 15-second and 30-second intervals led to an increase in response rate with time while a 2-minute interval showed a decline.

In another experiment using Mackworth's (1950) schedule ($\frac{3}{4}$, $\frac{3}{4}$, $1\frac{1}{2}$, 2, 2, 1, 5, 1, 1, 2, 3, and 10 minutes), the signal was transient, allowing measurement of both percentage detection and rate of observing over a 2-hour period. The similarities to Mackworth's results were good. Holland found 39% of his observers missed one or no signals, Mackworth found 29%. Separating the "good" observers from the "poor" observers, Holland plotted separate curves for the two groups relating percentage detected and rate of observing as a function of time in half-hour periods. The poor observers showed a sharp decline over the first half-hour in both measures, with a continued gradual drop until the last period where some recovery occurred. For the good observers percentage of signals detected did not decline and observing rate increased according to a negatively accelerated function of time.

Holland supplements his paper with analogies between findings of vigilance studies and animal studies which use the Skinnerian cumulative response frequency method of recording defended by Holland. He cited Brady (1956) as finding higher response rate for rats under Benzedrine, analogous to Mackworth's result (1950) of increased signal detection. Ferster and Skinner (1957) produced different rates of response using multiple reinforcement schedules in the same session with animal subjects and this corresponds to Nicely and Miller's (1957) result of higher probability of detection for the quadrant on a radar scope having a higher signal rate. Holland also discusses increased rate of responding following rest for both animals and vigilance studies. Holland further cites the evidence that higher room temperatures produce lower response rates in animals, quite analogous to Mackworth's finding (1950) that detection is lowered under these circumstances.

The method of studying vigilance behavior through observing responses is subject to criticism on several grounds. Requiring an overt response such as key pressing introduces an element into the situation that is not present in free scanning vigilance tasks. There is the implicit assumption not only that the viewer looks at the display every time he presses the key, but also that this is the same scanning response that would occur if the subjects were not required to press the key. An equally reasonable interpretation is that the subject presses the key rapidly in order to keep the display illuminated so he can scan when he wants to, and the very high rates of responding (Holland, 1957, 1958) suggest this as the case. Furthermore, repetitive rapid pressing of a key can produce work inhibition or fatigue. Changes in rate of responding under these circumstances would not necessarily reflect the same laws of observing if head and eye movements were to be measured directly and motor fatigue was trivial. A paper by Blair (1958) described an observing response that makes the correspondence to normal scanning more likely than in the case of key pressing, at least for the moving head component of visual scanning. The operator was in a darkened

room and had a continuous light source on his head that had to be directed at the display to see the critical signal. A light-sensitive germanium diode activated a recorder whenever the operator "looked" at the display, thereby giving a complete record of frequency and duration of observing responses. Only two of his five subjects exhibited Holland's finding of increased observing as signal time approached, suggesting difficulties that could be devastating for Holland's equivalence of Skinnerian observing responses and sense receptor orientations. Mackworth and Mackworth (1958) reported a precise method of measuring eye fixations with closed circuit television methods. The direct measurement of eye movements (head movements are excluded by the television method) under conditions where the observer is allowed to scan a display freely can be used to test the hypothesis that more remote responses such as pressing a key are equivalent and yield the same laws of observing. Until such verification is made the indirect approach to the study of vigilance must be viewed with some reservation. Perhaps this potential source of difficulty in developing laws for observing responses arose from applications of Wycoff's (1952) general definition that held the observing response to be that behavioral act which produces the discriminative stimuli correlated with reinforcement. Thus, orienting the eyes and head to receive stimuli being emitted from a display would be an example of an observing response, and Wycoff freely uses this example, but it is clear that his definition is broadly conceived and includes *any* response which produces the discriminative stimuli for the organism. Prokasy (1956) and Lutz and Perkins (1960) followed Wycoff's lead and studied observing responses which were not the sense receptor orientations so important in vigilance research. Holland (1957, 1958), however, proceeds one step further and interprets the observing response of key pressing to illuminate the display in a vigilance experiment as having direct correspondence with sense receptor orientations and to yield the same behavioral laws. This is an ad hoc assumption which can be, and must be, proved empirically in the laboratory.

Conclusions and Summary

The main shortcoming of our contemporary theories of vigilance would seem to be a casualness of formulation that makes the definitive testing of implications rather difficult. The inhibition hypothesis has an implicit organization provided by the classical conditioning paradigm, a wealth of relationships derived from the study of responses in the classical conditioning situation, and the several theoretical explanations of classical conditioning, but this framework for the vigilance problem appears to be more of an analogy than a scientifically useful theoretical system that is capable of rigorously accounting for the present facts and predicting new experimental findings. Even if we grant the classical conditioning schema a higher status than analogy, its capabilities for relating to the known data of vigilance experiments are limited. For example, Mackworth (1950) sees the period of continuous observation in a vigilance task as a period of experimental extinction. As this extinction period continues the expectation would be for a steadily decreasing probability of occurrence for the detection response. Yet, this typically is not the

case. Mackworth's own data showed the probability of detection function stabilizing at an intermediate level, with no apparent trend toward complete extinction, and Mackworth acknowledged this difficulty by suggesting that other factors such as expectancy and self-instructions might be required to account for the data trends. Additional explanatory shortcomings of classical conditioning are found in the high level of responsiveness induced by a high signal rate (the classical conditioning schema would predict just the opposite because high signal rate would constitute many massed extinction trials), the effects of intersignal interval, and the general failure of detection decrement to occur in complex tasks with multiple stimulus sources. With all of these weaknesses, it is doubtful whether the inhibition hypothesis deserves serious attention in any efforts directed toward developing a satisfactory theory. It is also doubtful whether Broadbent's attention hypothesis can be refined sufficiently to account for all of the known findings and to provide new deductions that can be given decisive experimental evaluation in the laboratory. Broadbent's loosely structured views generally have prevented them from being instruments to guide the experiments of laboratory workers in this area, and it is difficult to see them ever becoming useful unless a basic rephrasing of their tenets is undertaken to give the precision that a science asks of a theory.

The expectancy hypothesis cannot be said to have a precise expression but certainly it has been a good heuristic device in stimulating a number of experiments. Deese's initial expression of expectancy (1955) was a broad one and was based on the empirical relations between the response measure of detection probability on the one hand and independent task variables of signal rate and intersignal interval on the other. The principal research on expectancy however, has been by Baker (1958, 1959a, 1959b, 1959c) who has revised the Deese formulation of expectancy and in his research has mainly concentrated on the short range variations in performance as a function of intersignal interval rather than the overall detection level based on signal rate. While Deese was unable to verify his hypothesis about expectancy and the intersignal interval, Baker found some support for his elaborated version of expectancy which he related to more variables than Deese. Baker's expectancy hypothesis is quasiqualitative at this time but his approach is amenable to quantitative expression. It is not complete, however, because it does not try to account for decrement as a function of observation time, marked gains over rest, or the typical absence of decrements for multisource complex tasks.

The sensory variation or activationist hypothesis is secured in provocative physiological hypotheses about the role of ascending reticular activating system in maintaining responsiveness and, by appealing to a proposed organismic requirement for stimulus variation if performance level is to remain high, most of the facts of vigilance research can be explained in a general way. Behavior theories have always emphasized the guiding role of stimuli acquired through learning, and properly so, but these recent physiologically-based hypotheses stress that stimuli have a maintaining function for the response too. Thus, the monotony of the vigilance situation is interpreted to be an absence of stimulus variety needed to maintain the response level,

and variables such as high signal rate, rest, knowledge of results, task complexity, etc., are taken to be operations which promote stimulus variation and high responding. This activation hypothesis is useful after-the-fact but it remains to be expressed carefully before-the-fact so that differential prediction can be made. These clues derived from the physiological level are suggestive but they are insufficient by themselves. The theorist still has problems at the molar level where type and amount of external stimulation must be related to overt behavior. Ideally it is desirable to coordinate the molar concepts and functions with those at the physiological level, and the increased vigor of physiological research suggests that these echelons of organismic action eventually will be interrelated. But with all of our present vigilance data at the molar level, it would be fruitful at this time to look for an expression of the stimulus variation hypothesis in terms of stimulus control of responses for the whole organism. The possibilities for dimensions of stimulation, both on the external environmental and the response-induced side, seem endless. On the environmental side, the experimenter might systematically vary the number of stimulus sources, their spatial array, or the type and level of extraneous stimulation like noise or music. On the response side, the level of stimulation could be manipulated by variables influencing proprioceptive feedback, such as extent of movement or physical variables of the control system considered to be related to proprioception. Response-produced stimulation might also stem from mediating responses and be related to the number of choices involved in decision making. In the beginning a formulation of this kind could be a straightforward empirical expression in terms of stimulus and response without resort to intervening variables, but ultimately these variables would seem necessary for a sophisticated theory of vigilance behavior. As Hebb (1955) has suggested, the relationship between stimulus variation and responsiveness may be a motivational one.

REFERENCES

ADAMS, J. A. Vigilance in the detection of low-intensity visual stimuli. *J. exp. Psychol.*, 1956, **52**, 204–208.

BAKAN, P. Preliminary tests of vigilance for verbal materials. Memorandum Report No. B-2, March 31, 1952, University of Illinois.

BAKAN, P. Discrimination decrement as a function of direction of stimulus change. Memorandum Report No. H-1, June 15, 1953, University of Illinois.

BAKAN, P. Discrimination decrement as a function of time in prolonged vigil. *J. exp. Psychol.*, 1955, **50**, 387–390.

BAKAN, P. Extraversion-introversion and improvement in an auditory vigilance task. *Med. Res. Council, Appl. Psychol. Res. Unit Rep.*, 1957, No. APU 311/57.

BAKER, C. H. Attention to visual displays during a vigilance task: I. Biasing attention. *Brit. J. Psychol.*, 1958, **49**, 279–288.

BAKER, C. H. Attention to visual displays during a vigilance task: II. Maintaining the level of vigilance. *Brit. J. Psychol.*, 1959, **50**, 30–36. (a)

BAKER, C. H. Three minor studies of vigilance. *Def. Res. Med. Lab. Rep.*, 1959, No. 234-2. (b)

BAKER, C. H. Towards a theory of vigilance. *Canad. J. Psychol.*, 1959, **13**, 35–42. (c)

BARTLETT, SUSAN C., BEINERT, R. L., & GRAHAM, J. R. Study of visual fatigue and efficiency in radar observation. *USAF RADC tech. Rep.*, 1955, No. 55–100.

BLAIR, W. C. Measurement of observing responses in human monitoring. *Science*, 1958, **128**, 255–256.

BRADY, J. V. Assessment of drug effects on emotional behavior. *Science*, 1956, **123**, 1033–1034.

BROADBENT, D. E. The Twenty Dials Test under quiet conditions. *Med. Res. Council,*

Appl. Psychol. Res. Unit Rep., 1950, No. APU 130/50.

Broadbent, D. E. Classical conditioning and human watch-keeping. *Psychol. Rev.*, 1953, **60**, 331–339.

Broadbent, D. E. Some effects of noise on visual performance. *Quart. J. exp. Psychol.*, 1954, **6**, 1–5.

Broadbent, D. E. *Perception and communication.* New York: Pergamon, 1958.

Deese, J. Some problems in the theory of vigilance. *Psychol. Rev.*, 1955, **62**, 359–368.

Deese, J., & Ormond, E. Studies of detectability during continuous visual search. *USAF WADC tech. Rep.*, 1953, No. 53–8.

Ferster, C. B., & Skinner, B. F. *Schedules of reinforcement.* New York: Appleton-Century-Crofts, 1957.

Garvey, W. D., Henson, J. B., & Gulledge, Irene S. Effect of length of observing time on earth satellite visibility. *U. S. Naval Res. Lab. Rep.*, 1958(Feb), No. 5094.

Garvey, W. D., Taylor, F. V., & Newlin, E. P. The use of "artificial signals" to enhance monitoring performance. *U. S. Naval Res. Lab. Rep.*, 1959(Feb), No. 5269.

Harabedian, A., McGrath, J. J., & Buckner, D. N. The probability of signal detection in a vigilance task as a function of intersignal interval. Technical Report No. 3, February 1960, Contract Nonr 2649 (00), Office of Naval Research, Psychological Sciences Division, Personnel and Training Branch.

Hebb, D. O. Drives and the C.N.S. (conceptual nervous system). *Psychol. Rev.*, 1955, **62**, 243–254.

Hoffman, A. C., & Mead, L. C. The performance of trained subjects on a complex task of four hours duration. OSRD Report No. 1701, 1943, United States Department of Commerce.

Holland, J. G. Technique for behavioral analysis of human observing. *Science*, 1957, **125**, 348–350.

Holland, J. G. Human vigilance. *Science*, 1958, **128**, 61–67.

Howland, D. An investigation of the performance of the human monitor. *USAF WADC tech. Note*, 1958(Jul), No. 57–431.

Jenkins, H. M. The effect of signal-rate on performance in visual monitoring. *Amer. J. Psychol.*, 1958, **71**, 647–661.

Jerison, H. J. Experiments on vigilance: Duration of vigil and the decrement function. *USAF WADC tech. Rep.*, 1958, No. 58–369.

Jerison, H. J. Experiments on vigilance: The empirical model for human vigilance. *USAF WADC tech. Rep.*, 1959, No. 58–526.

Jerison, H. J., & Wallis, R. A. Experiments on vigilance: One-clock and three-clock monitoring, *USAF WADC tech. Rep.*, 1957(Apr), No. 57–206. (a)

Jerison, H. J., & Wallis, R. A. Experiments on vigilance: Performance on a simple vigilance task in noise and in quiet. *USAF WADC tech. Rep.*, 1957(Jun), No. 57–318. (b)

Jerison, H. J., & Wing, S. Differential effects of noise and fatigue on a complex vigilance task. *USAF WADC tech. Rep.*, 1957(Jan), No. 57–14.

Kappauf, W. E., & Powe, W. E. Performance decrement at an audio-visual checking task. *J. exp. Psychol.*, 1959, **57**, 49–56.

Krendel, E. S., & Wodinsky, J. Visual search in an unstructured visual field. *USAF Cambridge Res. Cent. Rep.*, 1959, No. 59–51.

Lindsley, D. B. Psychophysiology and motivation. In M. R. Jones (Ed.), *Nebraska symposium on motivation: 1957.* Lincoln: Univer. Nebraska Press, 1957. Pp. 44–105.

Loeb, M., & Jeantheau, G. The influence of noxious environmental stimuli on vigilance. *J. appl. Psychol.*, 1958, **42**, 47–49.

Lutz, R. E., & Perkins, C. C., Jr. A time variable in the acquistion of observing responses. *J. comp. physiol. Psychol.*, 1960, **53**, 180–182.

McCormack, P. D. Performance in a vigilance task as a function of inter-stimulus interval and interpolated rest. *Canad. J. Psychol.*, 1958, **12**, 242–246.

McFarland, R. A., Holway, A. N., & Hurvich, L. M. Studies of visual fatigue. Report, 1942, Harvard Graduate School of Business Adminstration.

Mackworth, J. F., & Mackworth, N. H. Eye fixations recorded on changing visual scenes by the television eye-marker. *J. Opt. Soc. Amer.*, 1958, **48**, 439–445.

Mackworth, N. H. The breakdown of vigilance during prolonged visual search. *Quart. J. exp. Psychol.*, 1948, **1**, 6–21.

Mackworth, N. H. Researches on the measurement of human performance. *Med. Res. Council spec. rep. Ser.*, 1950, No. 268.

Mackworth, N. H. Some factors affecting vigilance. *Advanc. Sci.*, 1957, **13**, 389–392.

Malmo, R. B. Activation: A neuropsychological dimension. *Psychol. Rev.*, 1959, **66**, 367–386.

Mowrer, O. H. Preparatory set: Some methods of measurement. *Psychol. Monogr.*, 1940, **52**(2, Whole No. 233).

NICELY, P. E., & MILLER, G. A. Some effects of unequal spatial distribution on the detectability of radar targets. *J. exp. Psychol.*, 1957, **53**, 195–198.

POLLACK, I., & KNAFF, P. R. Maintenance of alterness by a loud auditory signal. *J. Acoust. Soc. Amer.*, 1958, **30**, 1013–1016.

PROKASY, W. F., JR. The acquisition of observing response in the absence of differential external reinforcement. *J. comp. physiol. Psychol.*, 1956, **49**, 131–134.

SAMUELS, INA. Reticular mechanisms and behavior. *Psychol. Bull.*, 1959, **56**, 1–25.

SCOTT, T. H. Literature review of the intellectual effects of perceptual isolation. Report No. HR66, July 1957, Defence Research Board, Department of National Defence, Canada.

WYCOFF, L. B., JR. The role of observing responses in discrimination learning. *Psychol. Rev.*, 1952, **59**, 431–442.

(Received September 29, 1960)

Psychological Bulletin
1962, Vol. 59, No. 4, 273–283

MENTAL ABILITY AND SOCIOMETRIC STATUS AMONG RETARDED CHILDREN[1]

ROBERT A. DENTLER[2] AND BERNARD MACKLER
Bureau of Child Research, University of Kansas

As this review of recent research will indicate, there is continuing interest among students of mental retardation in the relationship between sociometric (or peer choice) status and the level of ability of the person being chosen. Our aim is to review and evaluate representative investigations of this relationship among normal children, institutional retarded children, and retarded children attending regular or special classes in public schools. But our focus is on the institutional retardate.

The educational and institutional importance of the relationship rests on the notion that the interpersonal environment is a powerful determinant of development, and on the further notion that the interpersonal environment of the child is composed predominantly of peer group relations. A given interpersonal environment may be assessed as facilitative or restrictive of development. Thus against school and institutional standards of training, achievement and performance, it is important to know the extent to which the peer environment rewards or penalizes members differentiated on abilities.

Like groups of adults, groups of children exhibit structures of differentiated preferences. Some members will be preferred more by others. Obviously, there are many correlates of such structures, and these vary for any group according to the criterion used for expressing preference. A preference structure will usually have as correlates measures of homophyly, homogamy, propinquity, social conformity, social initiative or dominance, as well as more esoteric aspects of personal attractiveness.

It is equally likely that the preference structure will reflect what is culturally valued within the group, as Riecken and Homans (1954) suggest. If academic achievement is valued by school children for example, the ablest peers (academically) will tend to be "overchosen." As other values may compete with achievement for peer endorsement, the relation between mental ability[3] and sociometric status will always be limited. The more intelligent children may also prove more competent in their expression or realization of competing values, however, which leads one to expect a persisting relationship between mental ability and sociometric status. This does *not* suggest that the correlation between ability and sociometric status should be high, but rather that it should be ubiquitous.

[1] This review was supported in part by United States Public Health Service Grant Number OM-111, and in part by a grant from the University of Kansas. We acknowledge the helpfulness of John de Jung, Gerald Siegel, and Ross Copeland in commenting critically on an earlier draft of this report.

[2] Now at Dartmouth College.

[3] Mental ability is used throughout this paper interchangeably with intelligence. All the studies reviewed have used measures of intellectual performance of the individual as against measures of "potential." In speaking of intellectual or mental ability, we hoped to avoid the many remedial, diagnostic, and other clinical connotations which accrue to the concept of the intelligence quotient. We are also aware, as our discussion indicates, of the need for a better conceptualization of mental ability.

NORMAL CHILDREN

In separate studies of five different samples of children in grades from second through seventh, Bonney (1944) and Laughlin (1954) found positive correlations between intelligence and sociometric status. In each case, the association was significant beyond the .01 level, but the coefficients (Pearson) were low, ranging from .31 ($N=299$) to .27 ($N=525$).

Grossman and Wrighter (1948) related intelligence as measured by the Stanford-Binet to sociometric status in a class of sixth grade students. They reported that high status peers were significantly higher in intelligence. They concluded that the two variables were significantly related, but that high intelligence did *not* assure high status.

Barbe (1954), Bonney (1946), and Potashin (1946) gave further confirmation to these findings, but they also found that mutual choices tended to be made between children with similar levels of mental ability. To this extent, the general association between ability and status is reduced insofar as children in the modal range of mental ability reflect higher sociometric status: there is a greater chance of being chosen for subjects of modal ability, and the range of status scores is restricted.

Gallagher and Crowder (1957), in a study of gifted children, found that four out of five students with Stanford-Binet IQs of 150 or above obtained above average sociometric status and that more than half scored in the top status quartile. Thus when relatively extreme cases are considered, mental ability improves considerably as a predictor of sociometric status.

Other representative or noteworthy studies of the relation among normal children are summarized in Table 1. When measures of specific types of performance such as reading comprehension, school achievement, motor skills, and social maturity have been correlated with sociometric status, coefficients of similar magnitude have been obtained. These findings are reviewed in Gronlund (1959).

RESEARCH ON INSTITUTIONAL RETARDATES

Using the choice of "one best friend" as a test criterion, Hays (1951) investigated the problem with a sample of 127 defective, borderline, and dull-normal girls housed in a single institutional dormitory. His subjects ranged in age from 7 to 23 years, with a mean of 14. Intelligence quotients and mental ages were derived from Stanford-Binet tests. The biserial correlation between dichotomized "choices received" and "no choices received" and IQ was .43 ($p<.01$).

Clampitt and Charles (1956) studied the relationship between sociometric status and supervisory evaluation of institutionalized mentally deficient children. The 164 subjects, both girls and boys, ranged in age from 6 to 40 years. The mean age for girls was 19, for boys 14. At the time of the study, all subjects had been institutionalized for at least 1 year, and the median term was 7 years for girls and 3 years for boys. Most of the intelligence tests were Stanford-Binets. Sociometric choice and rejection responses were obtained in relation to the following activities: eating, playing, and working. Significant, positive rank order correlations were found between sociometric status and MA, IQ, and supervisory evaluation based on selected traits. The correlation be-

TABLE 1

SUMMARY OF MAJOR STUDIES OF NORMAL CHILDREN

Authors	Choice criterion (or criteria) as obtained through sociometric test. (All choices within sample unless states otherwise.)	Measures of mental ability	*N*	Correlation coefficient and statistical test
Bonney (1944)[a]	Take a trip with, vote for class librarian, vote for class officers, best friend, take picture with, number of valentines received, have as a partner for Easter party, expect to give Christmas presents, and best citizens and best leaders. No limit was made on the number of choices.	California Test of Mental Maturity (grade 2), Kullman-Anderson (grades 3 & 4), Otis (grade 5), Pintner Intermediate (grade 6), Gates Primary Reading (grade 2), Stanford Achievement (grades 3, 4, 5, 6)	229	.31–.45 Pearson
Bonney (1946)	Same as Bonney (1944)	Same as Bonney (1944)	201	.28–.44 Pearson
Gallagher & Crowder (1957)	5 best friends	Stanford-Binet, WISC, Stanford Achievement Test	30	.45 contingency
Grossman & Wrighter (1948)	First three choices for: sit near, walk home, play, class officer, and best friend	Stanford-Binet, Stanford Achievement Test	117	"Usual low rectilinear relationship"
Laughlin (1954)	One of your best friends, choose a group member but not a close friend, choose someone to be with once in a while, don't mind this individual in group but do not want to have anything to do with him, and wish individual were not in the group	Detroit Alpha Intelligence Test, Metropolitan Achievement Test—partial battery	525	.27–.31 Pearson
Potashin (1946)[b]	First three choices for: favorite activity, classroom project, best friend in class, and best friend out of class	Dominion junior and intermediate group tests	124	
Rosenthal (1956)[c]	First three choices for: play, sit next to, invite to a party, and go to a show	Kuhlmann-Anderson language measures	358	

[a] Barbe (1954) in his study of 244 normal childern gives only percentages of intelligence levels of friends chosen by bright and slow learning children. No correlation test could be performed as distribution by IQ level was not given.

[b] Mean difference of friends (mutual choices) and nonfriends were compared as to mental age and intelligence. The difference in the MA and IQ of friends and nonfriends is 1.2 and .2, respectively; the reliability of the difference in the MA and IQ is .32 and .06, respectively.

[c] t tests compared high and low sociometric groups on 10 language measures. On six of the tests the groups differed significantly ($p \leq .05$), and on four of the tests nonsignificantly ($p \leq .10$).

TABLE 2

SUMMARY OF MAJOR STUDIES OF RETARDED CHILDREN

Authors	Choice criterion (or criteria) as obtained through sociometric test. (All choices within sample unless stated otherwise.)	Measures of mental ability	N	Correlation coefficient and statistical test
	Institutional retardates			
Clampitt & Charles (1956)	Three choices for: eating, playing, and working	Stanford-Binet	164	.34 Spearman
Dentler & Mackler (1960)	Three (or more) choices for: play, work, to be, and not want to play with	Porteus Maze	29	.50 Pearson
Farber & Marden (1958)	Three best friends	Stanford-Binet	77	.40 Spearman
Hays (1951)	One best friend	Stanford-Binet	127	.43 Biserial
McDaniel (1960)	One choice for: sit next to at lunch, sit with at movie, play, and work	WISC	15	.35 Spearman
Sutherland, Butler, Gibson, & Graham (1954)	Two choices for: working, eating, recreational periods, spare time activities, best friend, take with you on discharge, prefer to discuss plans and troubles, and associate with after marriage	Stanford-Binet	205	.34 Pearson
	Noninstitutional retardates			
Johnson (1950)[a]	Three choices for: like, sit next to, and play	Stanford-Binet (for retarded members, $N=39$), Vineland Social Maturity, New California Short-Form Test of Mental Maturity	688	
Turner (1958)[b]	Three choices for: sit next to and play	Otis Quick Scoring Mental Ability Test, Vineland Social Maturity	390	

[a] t tests compared typical group with mentally handicapped group as to acceptance ($t=4.10$, $p \leq .01$) and rejection ($t=4.94$, $p \leq .01$).

[b] t tests compared high chosen group with low chosen group on ability ($t=2.1$, $p \leq .05$) and social maturity ($t=2.78$, $p \leq .01$).

tween IQ and choice status for boys, for example, was .34 ($p<.01$).

Farber and Marden (1958) studied the social organization of a boy's unit at a state school for the mentally deficient. The sample contained boys ranging in age from 11 to 19. The mean age was 14.7. The length of residence ranged from 4 months to nearly 15 years. Only boys for whom Stanford-Binet IQ scores were available were included. For the 77 subjects, status ranks were ascertained by interviews and questionnaires. A rank correlation of .40 ($p=.01$) was found between IQ and status.

Sutherland, Butler, Gibson, and Graham (1954) made a sociometric study of one cottage of retarded females. The 205 subjects ranged in age from 18 to 53 years. The published report did not indicate the length of institutionalization for the cottage members. The Stanford-Binet (Form M) was used as the ability measure. For their subjects, Sutherland et al. found a Pearson coefficient of correlation of .34 ($p<.01$) between intelligence and choice status. To illustrate the relationship, Sutherland et al. compared high and low status subjects. The mean IQ of the high status group was 62.6 compared with a mean of 42.8 for the lows. Of the highs, 87% were above 70 IQ, while in the low status group 81% were below 70 IQ.

These four studies on institutional retarded children are corroborative. They compare point for point with studies of normal children and with one another in finding a positive, significant yet weak association between intelligence and sociometric status.

McDaniel (1960) studied 15 retardates, 3 women and 12 men, who ranged in age from 16 to 32 years, with a mean age of 19. At the time of the administration of the first sociometric test, the group had been in existence for 8 months. The study did not indicate the length of institutionalization, however. The mean IQ for the group, based on the Full Scale WISC test, was 52. The Spearman rank correlation between IQ and sociometric status was .35 ($p<.10$). Although this coefficient fails to differ significantly from zero at the .05 level, it is better to consider the coefficients under review in terms of magnitude and expected range. For example, it is likely that McDaniel would have found a correlation significant at p equal to or less than .05 had he assessed a group with 30 rather than 15 subjects; yet the actual magnitude of the coefficient would probably have fallen between .30 and .50.

McDaniel's subjects exhibited very little interaction and a restricted range of choices. Considering the "looseness" of this group's sociometric structure, the relation between mental ability and sociometric status McDaniel found is all the more indicative of the weak but pervasive character of differentiation by mental ability. Even where group position appears to have limited salience for members, social preference depends somewhat upon the demonstration of skills of value in group activity.

Dentler and Mackler (1960, 1961) investigated mental abilities in relation to sociometric status among 29 *newly arrived* boys in a state school for retarded children. The boys ranged in age from 6 to 12; mean IQ on the Porteus Maze Test was 56. After the first month of residence, sociometric and psychometric measures were taken. The association between mental ability[4] and sociometric status was .50 ($p<.01$,

[4] A full scale score was obtained from T scale scores on the Porteus, the Parsons Language Sample, and an index of social maturity.

Pearson *r*). In the second month, repeated sociometric assessment revealed a correlation between choice status and ability of −.14. Further analysis indicated that, under increased pressure from aides to restrict peer interaction and to induce conformity to cottage regulations, group structure was reorganized (at least temporarily) so that sociometric status became increasingly associated with conformity.

Research on Noninstitutional Retardates

Johnson (1950) conducted a study in two communities in which there were no special classes for the mentally retarded, thus assuring that all the educable, mentally retarded were in regular classrooms. Grades 1 through 5 were sampled. He found that children diagnosed as mentally handicapped obtained significantly lower sociometric status scores than the nonhandicapped. In addition, Johnson found that sociometric status was directly related to IQ and that rejection scores were inversely related to IQ.

Turner (1958) studied the sociometric status of mentally retarded children enrolled in special classes in Negro elementary schools in North Carolina. In all, 18 classes were sampled and 390 children tested. Using three measures to assess mental ability (Table 2), Turner found that high ability children were chosen 15 or more times and the lows 3 times or less, using roughly the top fourth and the bottom fourth of the subjects ranked on mental ability.

Measures of Sociometric Status

The techniques and choice criteria used to measure sociometric status in the papers reviewed vary greatly. Barbe (1954) had teachers ask their pupils to nominate their three best friends, and gave equal weight to each choice. Gallagher and Crowder (1957) asked for nomination of five best friends and ranked subjects by number of choices received. Bonney (1944, 1946) in marked contrast, used event-specific criteria. His subjects chose peers with whom they would most like to have their pictures taken, those they would prefer to work with on a committee for a social event, have as partners for a trip to a packing house, and so forth, across three additional criteria. Moreover, number of valentines received on Valentine's Day was tabulated and included as an indicator. Criteria were varied from class level to level. Scores were weighted, and number of choices was not limited. Resulting frequencies of choices received were calculated as proportions of totals.

Grossman and Wrighter (1948) used 10 choice criteria. They were thus able to assess internal reliability and found a mean Spearman-Brown reliability coefficient for four samples of .95. Validity was checked by examining the fit between sociometric status and children elected as class officers. Scores were weighted and number of choices received on each criterion were summed.

In studies involving institutional retarded children, measures must be fitted to surmount illiteracy as well as other handicaps. Clampitt and Charles (1956) used three choice criteria: eating associates, playmates, and workmates. Choices were elicited in interviews, and probes were used to clarify communication and to insure at least three choices on each criterion. Number of choices was across the three criteria, with rejections being weighted negatively. Farber and Marden (1958) interviewed their subjects and elicited unlimited nominations of best friends. If a subject

named only persons not included in the sample, he was asked to name his best friends within the group. First, second, and third choices were weighted. Hays (1951) also used individual interviews but asked only for choice of one best friend. Number of choices received ranged from zero to eight. Biserial correlation was necessary because of the extremely skewed distribution. McDaniel (1960) interviewed his institutional subjects but employed six criteria, including preferred associates to lunch with, sit with at the movies, play with, work with, help on a job, and persons nominated as those with whom the subject would not do any of these things. Only one choice was elicited on the first five criteria. Subjects were ranked by total number of choices received. Interestingly, no rejection nominations were made. McDaniel retested the group and obtained a response stability coefficient of .60 (Spearman). Sutherland and associates (1954) also interviewed their institutional subjects, eliciting two choices on each of eight criteria. Most of these were identical to those used by McDaniel but best friends were nominated as well as choices of associates preferred after institutional release. Of all the studies of institutional subjects, only Sutherland's employed a probability model (Bronfenbrenner, 1945) to categorize subjects by status level.

Dentler and Mackler (1960) attempted to simplify choice elicitation by using photographs of all group subjects randomly arrayed in rows and columns on a large but portable beaverboard. Children were interviewed under informal conditions and asked to point to their choices on four criteria: playmates, workmates, most want to "be," and not want to play with. Number of choices received per criterion was normalized and resulting scores were combined.

In their studies of sociometric status of noninstitutional retarded children, Johnson (1950) and Turner (1958) asked for friendship nominations, playmates, and classroom seatmates. Both obtained their data through personal interviews but only Johnson secured rejection choices. Again, choices were not weighted, and summed number of choices served as the scale.

Discussion

This review suggests that a moratorium could be declared on studies of the relation between intelligence and sociometric status among children—a moratorium that ought to hold whether the children are gifted, normal, or mentally retarded, and whether or not they are institutionalized. The relationship has been demonstrated to hold, and to hold at a characteristic level, on samples ranging in size from 15 to more than 500, and across a wide age range. Its strength tends toward a constant value even where group relations have not become well established. Among retarded children in institutions, length of institutional residence appears to have little effect on the general relationship. The coefficients hold whether the intelligence test is the Stanford-Binet, the WISC, the California test, or the Porteus Maze, and whether it is individually or group administered. Finally, the relation is roughly the same whether IQ or mental age is used.

Most of the studies employed the best available instruments for assessing school related aspects of intelligence. In research involving retarded children, however, it is important to exploit the fact that a large variety of abilities exist and that some of these are probably more

closely related to sociometric status than others. For example, Rosenthal (1956) found that the *language* of children of high sociometric status was more active and moving and more varied.

It should be possible to differentiate relations between a variety of abilities *and* a variety of types of sociometric statuses. Thus functioning abilties such as performance subscales on intelligence tests or motor skills might be very highly associated with sociometric status on criteria involving leisure association or playmate preferences. To avoid premature closure of the question of how abilities relate to group status, at this stage it may be much more useful to consider discrete performance measures and discrete status indicators.

These considerations regarding the "true" range of group related abilities apply also to studies of normal children in classroom situations. As Gronlund (1959) indicates, research on measures of ability in relation to sociometric status among normals has accumulated steadily since 1940, yet little has been done to develop measures of the kinds of abilities that might be assumed, on the basis of hypotheses about group structure, to have peculiar relevance as determinants of status. Most projects have employed instruments developed originally by educational psychologists and clinicians for very different purposes.

Some factors underlying the *low* correlation between mental ability and status are methodological; others are substantive. Only one of the studies reviewed made use of a probability model for classifying students as high, medium, or low in status, suggesting that much of the presumed differentiation between subjects may be due to little more than chance or error variance. Similarly, only one study employed methods of computation which took into account the ones making choices as well as how many choices were received. Heavy reciprocation in the modal range could distort and depress estimates of true association. Only one study assessed in detail the interrelations between results on the several choice criteria. A few more undertook evaluation of the reliability of the responses, though unfortunately, three of the studies treated repeated measurements of sociometric status as tests of reliability rather than as indicators of change. Users of sociometric measures should accept the probability of change over time. The task of specifying the elements of change in test data that may be attributed to the actual changes in the variable under study, as opposed to change that must be attributed to unwanted or chance flunctuation in the test, is a task that has not been undertaken in the research under review.

Despite the wide range of work attesting the reliability and validity of diverse measures of sociometric status (Mouton, Blake, & Fruchter, 1960a, 1960b), there is little doubt about the need for clarification of the concept of sociometric status. There is agreement that sociometric status gives an indication of the differential value peers tend to place on each other. There is also agreement that groups have standards against which such valuation is made. To this extent, clarification by empirical means should be possible through closer attention to these norms or standards. There is no good basis for demanding that choice criteria should be highly intercorrelated or even highly stable over time, but there is theoretical basis for investigating the fit between

criteria and group standards. For example, none of the studies reviewed found length of residence in the institution to be a qualifying variable, yet one of the concerns of the sociometrist should be the analysis of "institutional effects." How do groups of children within institutions develop structure, and how are these structures influenced by the crucial fact of their location within an institutional culture?

Length of residence may be determinative if approached longitudinally. Group structures are *emergents;* thus, sociometric status within an institutional cottage which has just formed may differ greatly from status in a cottage that has endured for years. The problem is one not of length of residence of individuals perhaps, but of duration of the group, as McDaniel (1960), Farber and Marden (1958), and Dentler and Mackler (1960) suggest.

The study by Farber and Marden (1958) points a path that should be followed. Beyond finding association between intelligence and status, these investigators demonstrated associations between sociometric status and formal classification of institutional boys as educables, trainables, and working-boys; status and popularity; and status and history of delinquency. They identified the bases on which institutional retarded children in at least one state school classify themselves. For example, their subjects distinguished between peers oriented to rehabilitation and those disposed toward a "custodial career." Sociometric status was shown to be associated with such classifications. So used, status serves as a key to understanding the career paths provided by the institution and chosen by the patients. Though their study does not treat changes in status, Farber and Marden have developed means for predicting the future behavior of institutional retarded boys.

Dentler and Mackler (1961), in studying *changes* in sociometric status among newly arrived patients, found that as the culture of the institution was absorbed, the relation between mental ability and status came to depend increasingly upon standards imposed on the group by aides. For at least a brief period, the usual positive relation between mental age and sociometric status was reversed. The mentally abler boys resisted the regulations most strongly and lost status as a result of deviance.

Future sociometric research on the differentiation of members within children's groups should specify with greater precision the nature of the performance or ability under assessment and the particular variety of status. This effort should be linked with collection of data relevant to the development, situation, and normative content of the group. Global or general indicators should be abandoned.

Summary

A review of representative studies of the relation between ability and sociometric status among normal children, institutional mentally retarded, and noninstitutional retarded children, indicated high agreement with the generalization that individual ability is positively and significantly associated with choice status. Studies of normal children have demonstrated that this relation holds whether the abilities assessed are measures of mental age, intelligence quotients, or quite different measures of achievement, or social or motor skills. Although significant, the association is uniformly limited to the .25 to .50 range.

Sociometric status studies of institutional retarded children were viewed as particularly important, as they provide access to the study of institutional effects and practical evaluations of social rehabilitation. Sociometric research on institutional children has been limited to correlation analysis of relations between sociometric status and school-type intelligence tests, length of residence, and age, with the exception of but a few reports.

The reviewers proposed that future studies should sharpen the concept of mental ability or include dimensions that concepts of group structure suggests are of probable importance in a given situation. Studies that attend exclusively to the relation between intelligence and status should be avoided, while efforts to predict status within groups undergoing formation or change should be increased.

REFERENCES

Barbe, W. B. Peer relationships of children of different intelligence levels. *Sch. Soc.*, 1954, **80**, 60–62.

Bonney, M. E. Relationships between social success, family size, socioeconomic home background, and intelligence among school children in grades III to V. *Sociometry*, 1944, **7**, 26–39.

Bonney, M. E. A sociometric study of the relationship of some factors to mutual friendship on the elementary, secondary, and college levels. *Sociometry*, 1946, **9**, 21–47.

Bronfenbrenner, U. The measurement of sociometric status: Structure and development. *Sociom. Monogr.*, 1945, No. 6.

Clampitt, R. R., & Charles, D. C. Sociometric status and supervisory evaluation of institutionalized mentally deficient children. *J. soc. Psychol.*, 1956, **44**, 223–231.

Davis, J. A. Correlates of social status among peers. *J. educ. Res.*, 1957, **50**, 567.

Dentler, R. A., & Mackler, B. Effects on sociometric status of institutional pressure to adjust among retarded children. Unpublished manuscript, University of Kansas, Bureau of Child Research, 1960.

Dentler, R. A., & Mackler, B. The socialization of institutional retarded children. *J. Hlth. hum. Behav.*, 1961, **2**, 243–252.

Dentler, R. A., & Mackler, B. The Porteus Maze Test as a predictor of functioning abilities of retarded children. *J. consult. Psychol.*, 1962, **26**, 50–55.

Farber, B., & Marden, P. High-brows and low-grades on boy's ward: Report of a study of the social organization of a boy's unit at a state school for the mentally deficient. Unpublished manuscript, University of Illinois, Institute for Research in Exceptional Children, 1958.

Gallagher, J. J., & Crowder, T. The adjustment of gifted children in the regular classroom. *Except. Child.*, 1957, **23**, 306–312, 317–319.

Gronlund, N. E. *Sociometry in the classroom.* New York: Harper, 1959.

Grossman, Beverly, & Wrighter, Joyce. The relationship between selection-rejection and intelligence, social status, and personality amongst sixth-grade children. *Sociometry*, 1948, **11**, 346–355.

Hays, W. Mental level and friend selection among institutionalized defective girls. *Amer. J. ment. Defic.*, 1951, **56**, 198–203.

Johnson, O. A study of the social position of mentally handicapped children in the regular grades. *Amer. J. ment. Defic.*, 1950, **55**, 60–89.

Laughlin, F. *The peer status of sixth and seventh grade children.* New York: Teachers Coll., Columbia Univer., Bureau of Publications, 1954.

Lindzey, G., & Borgatta, E. F. Sociometric measurement. In G. Lindzey (Ed.), *Handbook of social psychology.* Vol. 1. *Theory and method.* Cambridge, Mass.: Addison-Wesley, 1954. Pp. 405–448.

McDaniel, J. Group action in the rehabilitation of the mentally retarded. *Group Psychother.*, 1960, **13**, 5–14.

Mouton, Jane S., Bell, R. L., Jr., & Blake, R. R. Role playing skill and sociometric peer status. In J. L. Moreno (Ed.), *The sociometry reader.* Glencoe: Free Press, 1960. Pp. 388–398.

Mouton, Jane S., Blake, R. R., & Fruchter, B. The realiability of sociometric measures. In J. L. Moreno (Ed.), *The sociometry reader.* Glenocoe: Free Press, 1960. Pp. 320–361. (a)

Mouton, Jane S., Blake, R. R., & Fruch-

TER, B. The validity of sociometric measures. In J. L. Moreno (Ed.), *The sociometry reader*. Glencoe: Free Press, 1960. Pp. 362–387. (b)

POTASHIN, REVA. A sociometric study of children's friendships. *Sociometry*, 1946, 9, 48–70.

RIECKEN, H., & HOMANS, G. Psychological aspects of social structure. In G. Lindzey (Ed.), *Handbook of social psychology*. Vol. 2. *Special fields and applications*. Cambridge, Mass.: Addison-Wesley, 1954. Pp. 786–832.

ROSENTHAL, F. Some relationships between sociometric position and language structure of young children. Unpublished doctoral dissertation, University of California, 1956.

SUTHERLAND, J. S., BUTLER, A. J., GIBSON, D., & GRAHAM, D. M. A scoiometric study of institutionalized mental defectives. *Amer. J. ment. Defic.*, 1954, 59, 266–271.

TURNER, MILDRED W. A comparison of the social status of mentally retarded children enrolled in special classes. Unpublished doctoral dissertation, University of Indiana, 1958.

(Received January 19, 1961)

Psychological Bulletin
1962, Vol. 59, No. 4, 284–295

RESPONSE STYLE AS A PERSONALITY VARIABLE: BY WHAT CRITERION?

RICHARD K. McGEE[1]
Moccasin Bend Psychiatric Hospital, Chattanooga, Tennessee

Without doubt, one of the most active research areas in psychology during the last decade has been the study of test taking response sets, or styles. In particular, attention has been devoted to investigating their influence on personality inventory scores. This work has been confined almost exclusively to only three types of response tendency: the *social desirability* set, characterized by the consistent endorsement of desirable traits and the denial of undesirable ones; the *deviation* of a pattern of scores from the typical pattern produced by a given population of responders; and the *acquiescence* set, which consists of tendencies to choose the "true," "agree," or "like" option rather than their respective negative alternatives. Jackson and Messick (1958) have reviewed the research in this area, and have outlined their own suggestions for directing the course of future investigations. The purpose of the present review is to discuss one of the specific trends which has appeared in the area, subsequent to, and perhaps largely attributable to the Jackson and Messick (1958) article.

The development of particular interest is in the utility of the response set component of test scores. Whereas Lentz (1938) and Cronbach (1946, 1950) urged the control or elimination of noncontent determined variance, Jackson and Messick (1958) suggest that "for certain purposes in personality assessment opportunities for the expression of personal modes for responding should be enhanced and capitalized upon" (p. 244). Thus, the recent trend referred to above is based on the thesis that a response style has its roots in the underlying personality complex of the responder. It is proposed that individuals who vary in the extent to which they manifest a particular style of responding, will also vary in terms of certain measurable personality traits. Various dimensions of personality have been suggested, and evidence collected to support this hypothesis. The most recent and most provocative article in this series (Couch & Keniston, 1960) concludes on the assertion that

> this integrated study . . . has demonstrated both the far-reaching importance of response set in the area of psychological tests and the major proposition that the agreeing response tendency is based on a central personality syndrome (p. 173).

The relationship between response styles and personality traits appears to be a most promising problem for investigations in the near future. However, it is clear, even at this early stage, that the already available literature offers important implications for the design and execution of future studies. It is to this question that the present review is addressed.

Personality Correlates of the Social Desirability Response Style

No effort will be made here to review the vast literature which has

[1] The author is indebted to Douglas Jackson and Lee Sechrest for reviewing the original manuscript and offering their valuable suggestions for the final draft. Appreciation is also expressed to H. J. Wahler for initially stimulating the author's interest in response style research.

accumulated on this topic since Edwards (1953) reported a correlation of .87 between the scaled social desirability of item content and the frequency with which it is endorsed. In general, research with this response style has continued to be directed at showing its influence on a variety of psychological inventories (Bendig, 1959; Cowen & Tongas, 1959; Taylor, 1959, 1961), or its appearance in various clinical groups (Wahler, 1958). Recently research designed to study the complexity of the social desirability response style has revealed its multidimensional character (Messick, 1960a) and its interaction with other response styles, particularly acquiescence (Jackson, 1960; Jackson & Messick, in press; Messick, 1960b; Messick & Jackson, 1961).

The present literature contains only two suggestions that social desirability responding is related to basic personality traits. The first of these is barely more than a tentative guess. Allison and Hunt (1959) investigated the relationship between social desirability responding and the expression of aggression with various degrees of frustration. Social desirability tendency was measured by the Edwards Social Desirability scale (*SD* scale) and correlated with expression-of-aggression scores from a paper-and-pencil situational frustration test. In two experiments, they found that high *SD* scale scores are associated with a suppression of aggression in ambiguous situations where the culturally acceptable response is unspecified. Their tentative conclusion was that high social desirability tendencies are found in subjects who are "other-directed," whereas low social desirability tendencies characterize "inner-directed" individuals.

Crowne and Marlowe (1960) criticized the usual approach to the social desirability response style. They point out that one is never certain when to invoke the more parsimonious explanation that denial of undesirable traits is due to the genuine absence of the psychiatric symptoms usually embodied in the self-description inventories supposedly most affected by this tendency. Similarly, endorsement of desirable traits may reflect either defensiveness, or candid self-appraisal, especially in college students.[2] Hence, they offer a substitute method for measuring the social desirability tendency. The Marlowe-Crowne Social Desirability (M-C *SD*) scale is a 33-item inventory. Like the MMPI *L* scale, these items suggest behaviors which while socially desirable, cannot be endorsed by most people, if they are answering truthfully. It is the authors' contention that social desirability responding rests on a basic need of the individual to be accepted and approved of socially.

To test this notion, Marlowe and Crowne (1961) studied the relationship between *SD* scores and behavioral tasks in the laboratory. They employed the Spool Packing task developed by Festinger and Carlsmith (1959). This is a boring, seemingly meaningless task designed to arouse negative, antagonistic feelings. They predicted that individuals with high *SD* scores have a strong need for social approval, and

[2] This point is made by Crowne and Marlowe with specific reference to responses made by college students serving as subjects in social desirability research projects. It is not necessarily a criticism of the usual clinical interpretations of social desirability scores, such as Wahler's (1958) prognostic index for psychotherapy candidates, or the various MMPI scales designed to assess defensiveness. However, two recent studies (Jackson & Messick, in press; Messick & Jackson, 1961) have demonstrated the response set influence in the MMPI, and discussed important implications for the validity of these scales as they are currently used clinically.

would thus hold favorable attitudes toward the experimental situation following the Spool Packing task. More negative attitudes were predicted for the low *SD* scorers. *SD* scale scores were dichotomized at the mean, and these two groups were shown to differ significantly ($p<.01$) on the attitudes they expressed toward the task. The difference was in the predicted direction, with the high *SD* group expressing more favorable attitudes than the low *SD* group. They also observed a correlation of $-.54$ between the M-C *SD* scale and the Barron Independence of Judgment scale (1953) which is designed to measure social conformity.

To follow up this latter finding, Strickland (1960) administered the *SD* scale to subjects and also observed their behavior in an actual conformity situation similar to the original Asch (1956) procedure. When the *SD* scores were again dichotomized at the mean, the groups differed significantly ($p<.005$) on the basis of their yielding scores, with yielders having the higher need for social approval.

The preceding data are not presented here for the purpose of showing the construct validity of the M-C *SD* scale. These studies are important in that they represent a major attempt to relate social desirability responding to underlying personality variables. It is even more important that these studies have employed a procedure which is rare in the area of response style research. Seldom does one find studies wherein the stylistic variable is correlated with methodologically independent observations. The typical procedure has been to employ as criteria, other psychometric instruments containing a possibly strong methodological contamination. This attempt to seek an independent criterion is a highly valued step in psychological research, and is a point on which this review will focus later.

Personality Correlates of Deviant Response Patterns

Several studies have appeared in recent years dealing with general aspects of personality which relate to deviant responding in a variety of stimulus situations. This research has usually been presented as an attempt to validate the Deviation Hypothesis. Berg (1955, 1957, 1958, 1959) has formulated the Deviation Hypothesis notion as an extension of the concept of "set" or *Einstellung*. He suggests (1959) that it serves as a unifying principle to account for the results of many disparate studies seeking to predict behavior under widely varying conditions. In simplest form the Deviation Hypothesis asserts that deviant behavior is general; that deviant responses occurring in one "uncritical" area of behavior predict the occurrence of deviant responses in other "critical" areas. In a recent paper Sechrest and Jackson (1960) pointed out the broad generality of Berg's notion.

> It has been suggested that psychotics, lawyers, cardiac patients, transvestites, young normal children, character disorders, the obese, the feeble minded, psychoneurotics, and persons suffering from constipation, among others, represent deviant groups which might be expected to manifest their particular propensities toward deviation not only in a modality relevant to their particular symptoms and to items with relevant content, but also in response to one or more of the following: preference for abstract drawings, food aversion questionnaires, stimuli for conditioned responses, autokinetic and spiral aftereffect situations, vocabulary test items, figure drawings, musical sounds, and olfactory stimuli (p. 2).

Evidence to support this position has been presented by Grigg and Thorpe (1960). They administered the Gough 300-item adjective check-

list to a college freshman class and computed the frequency with which each item was checked by students as being self-descriptive. Those adjectives endorsed by more than 86% and those checked by fewer than 14% of the students were selected and formed a final 72-item list. This adjective checklist was then presented to the next incoming freshman class, and their self-descriptions obtained. Deviation scores were computed for each student by counting the number of commonly checked adjectives which he omitted, plus the number of rarely endorsed items he checked. At the end of the academic year students in this sample who appeared at the counseling center for vocational guidance, or for personal counseling, or who sought private psychiatric care in the community were identified. A control group was randomly selected from among the freshmen who did not fall in any of these three categories. When the deviation scores were compared for these four groups it was found that those seeking either private psychiatric care or personal counseling for emotional problems had significantly higher scores ($p<.01$) than those who sought only vocational guidance, or no help at all. The two higher groups were not significantly different from one another, nor were the two lower groups.

This study is the most recent investigation of the type initiated by Berg and Collier (1953). They demonstrated that the tendency to give extreme responses was a deviant pattern which would differentiate high anxiety males from low anxiety males. They used the ambiguous pictures of the Perceptual Reaction Test (PRT) (Berg, Hunt, & Barnes, 1949) as a response set measure. Barnes (1955) demonstrated that psychiatric patients with various diagnostic labels could be differentiated from one another, and from normal control subjects on the basis of their pattern of responses to the PRT.

It should be evident that the studies reviewed here are dissimilar in one major respect to others in this area; namely, there is an absence of any attempt to specify a *particular* trait or personality variable which is related to deviant responding, and therefore considered basic to it. While Berg has clearly shown that different personalities differ in the response pattern they produce, he has apparently not considered it necessary to hypothesize an underlying personality trait, and show how it relates to deviant responding. However, Berg (1959) has acknowledged several unresolved questions in his unfinished work. He suggests that the forthcoming research may be expected to provide important data concerning the environmental, personality, and biochemical variables which may be related to atypical response patterns.

Sechrest and Jackson (1960) consider the deviation of response patterns from group to group to be an extremely important research area with far-reaching implications for personality assessment. Yet, having learned from the "school of hard knocks and tough breaks" that psychological processes are not always simple unidimensional variables, they voice a healthy skepticism that the Hypothesis is really as general as Berg would apparently have his audience believe. Fortunately, Berg (1959) identified, as one of the major sources of difficulty in his work, the lack of operationally clean criteria for the identification of deviant and criterion groups. He has stressed the importance of selecting these groups on the basis of *valid behavioral characteristics*. It is to be expected that

future studies will attempt to show via operational criteria the extent to which the Deviation Hypothesis is applicable to measuring specific personality traits, and the conditions under which its generality is limited.

PERSONALITY CORRELATES OF RESPONSE ACQUIESCENCE

The tendency to respond "yes," "agree," or "true" to personality inventory items irrespective of their content has been the subject of many studies in recent years. In reviewing this area, Jackson and Messick (1958) concluded:

> In the light of accumulating evidence it seems likely that the *major common factors in personality inventories of the true-false or agree-disagree type*, such as the MMPI and the California Psychological Inventory, *are interpretable primarily in terms of style rather than specific item content* (p. 247).

In line with the extensive interest in response acquiescence per se, there have been numerous suggestions that herein lies a new device for observing systematic behavior which will lead to valid inferences about the nature of a particular "black box."

Authoritarians and Conformers

The acquiescence set first attracted major attention in connection with its influence on the California F Scale (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950). Thus it is quite natural that acquiescence has been closely linked with the trait of authoritarianism. When it was no longer reinforcing to crucify the F Scale because of its susceptibility to response style influence (Bass, 1955, 1957; Chapman & Campbell, 1957; Cohn, 1953; Jackson & Messick, 1957; Jackson, Messick, & Solley, 1957; Messick & Jackson, 1957) attention turned to the question of a psychological (as well as mechanical) relationship between response acquiescence and authoritarianism. Leavitt, Hax, and Roche (1955) suggested that the confounding of acquiescence and authoritarianism in the F Scale was a lucky accident which increased the discriminating power of the instrument. This conclusion was based on their view that the tendency to agree with things said in an authoritative manner is itself a factor of the authoritarian personality. Gage and his associates (Gage & Chatterjee, 1960; Gage, Leavitt, & Stone, 1957) have argued that negative items have more validity for measuring authoritarianism than do positive items. Their reasoning rests on certain assumptions about response acquiescence. They point out that disagreeing requires more self-confidence, ego strength, and personal security than does the act of agreeing. Hence, acquiescence is one of a family of traits including authoritarianism, conformity, or obeisance to authority. The person who responds in an acquiescent manner essentially yields to the "authority" of the printed word, or the physical stimulus however presented.

In a series of research projects Jackson (1955, 1958, 1959) has accumulated data which tend to confirm those "logical" arguments made by Gage cited above. In his earlier studies Jackson (1955, 1958) formulated a theory of cognitive energy. This hypothetical construct is inferred from a person's ability to resist field forces presented by stimuli in his environment. The number of perceptual shifts made by reversible figures under instructions to hold one phase represents an operational measure of a subject's resistance to hypothetical forces in the perceptual field. Jackson demonstrated that this measure is positively correlated with social conformity. Individuals who are able to "hold" the Necker cube in the

position instructed also have high independence scores on the Independence of Judgment scale; subjects who have less resistance to the tendency to perceive changes in the position of the cube are also yielders, or conformers. Recently, Jackson (1959) has shown that resistance to these field forces is associated with acquiescence response tendency as measured by F Scale scores. High acquiescers are low in cognitive energy whereas nonacquiescers are high in the energy required to resist the reversing of the figures. Thus, Jackson has presented empirical evidence to show that acquiescers are both conformers, and possessers of limited personality strength or energy. In so doing, he confirmed two of the predictions made by Gage and his associates, referred to above. Certainly his data are more convincing than those which led Bass (1956) to the tentative conclusion that the person with a high social acquiescence score is "an 'outward-oriented,' insensitive, non-intellectual, socially uncritical individual; in short, a Babbitt—an unquestioning conformer to social demands placed upon him" (p. 297).

Noncritical Thinkers

While Jackson has shown that acquiescence relates to a general process of cognitive functioning, specific reference has been made to the relationship between response style and critical, or analytical thinking. Frederiksen and Messick (1958) employed Helmstadter's (1957) method of separating the content component from the response set component of test scores. They observed the variance due to response set in relationship to nine of the personality scales of the Personality Research Inventory (Saunders, 1955). In general they found low correlations, but concluded that their data suggested the possibility of using response sets to measure some personality variables. Of particular interest in their report is the attention given to the trait of "criticalness," defined in terms of tasks employed in the study. It was shown that a set to be critical could be effectively induced for some tasks, but also that a significant (albeit low) negative correlation existed between criticalness and acquiescence measured by the F Scale. Hence, this would tend to corroborate Bass' (1956) assertion regarding the uncritical acceptance of situations by acquiescers. It might also be taken as confirmation of Jackson's notion relating cognitive energy level to nonacquiescence, assuming that critical or analytical thinking requires the exertion of a relatively high level of effort.

Additional data showing the interaction between acquiescence and certain cognitive variables have been published by Messick and Frederiksen (1959). They showed negative relationships between acquiescence to the F Scale (both original and reversed forms) and verbal knowledge, general reasoning, and deductive thinking. Previously Hardy (1956) had shown that certain scales of the CPI significantly predict academic achievement in a midwestern college population. Jackson (1960) observed that a feature which these scales had in common was a large number of items keyed "false," particularly items for which a "true" response would have been undesirable. Combining these separate observations, Jackson and Pacine (1960) reasoned that an acquiescence style, moderated by item desirability, should have a relationship to academic achievement. They examined this hypothesis and found that a criticalness style did predict grade-point averages to a low but significant degree. Acquies-

cence scores on modified F Scale forms did not predict academic achievement, but showed significant negative correlations with verbal knowledge, and consistent (not always significant) negative relationships with general reasoning.

There appear to be rapidly expanding pools of data which indicate that there are stable and meaningful relationships between the response determinants stimulated by true-false or agree-disagree item forms and measures of important cognitive variables.

Yeasayers and Naysayers

Perhaps the most ambitious research on the personality correlates of the acquiescence style has been presented by Couch and Keniston (1960). They combined 681 items from several personality inventories. A factor analysis of responses to these items yielded a 360-item agreement factor which the authors labeled the Overall Agreement Scale (OAS). With additional measures, they found positive correlations between OAS and scales with a high proportion of responses keyed true. (Where the greater proportion of items was keyed false, the correlations with OAS were negative.) Traitwise, high OAS was associated with measures of impulsivity, dependency, anxiety, mania, anal resentment, and anal preoccupation; low OAS was associated with ego strength, stability, responsibility, tolerance, and impulse control.

In addition to correlating OAS with these paper-and-pencil measures the authors made a searching clinical evaluation of their extreme responders. Subjects were selected from each tail of the distribution of OAS scores; high scorers were identified as yeasayers, and low scorers were labeled naysayers. Each subject then filled out a 55-item incomplete sentence form and participated in a depth interview lasting from 2 to 4 hours during which time the experimenter focused on each of these 55 projective responses. Following the interview, the experimenter rated each response on five separate scales, indicating the extent to which the response was typical of the theoretical yeasayer, or the theoretical naysayer. Interviews were "blind" with respect to knowledge of the subject's OAS score, and subjects were randomly divided among interviewers. Results revealed clear differences between the ratings made for yeasayers and naysayers. The authors describe these differences using the typical abstract clinical language: yeasayers are impulsive, emotionally reactive, extraverted, externally oriented, low in psychological inertia, and possess passive egos; naysayers are guarded, defensive, constricted, inhibited, introverted, withdrawing, introspective, high in psychological inertia, slow and critical reactors, and possess active egos. In summarizing their report, the authors consider the dimension of *Stimulus Acceptance* versus *Stimulus Rejection* to be the best single construct subsuming all the other specific traits related to agreeing response style. The similarity of this position to that of Bass (1956), Frederiksen and Messick (1958), and Jackson (1959) is quite significant.

Webster (1960) utilized the Couch and Keniston (1960) stimulus acceptance-rejection concept with his own speculation that response set (RS) variance is related to an ***inhibition*** versus ***lack of inhibition*** dimension. He concluded that another "all-pervasive syndrome" is being isolated for the understanding of personality. This conclusion was based on Webster's data which showed that RS had high negative correlations with measures of Schizoid Functioning and Impulse Expression. His RS

scores are determined by the frequency with which the subjects respond "no" so as to deny undesirable traits implying psychopathology. In line with the Crowne and Marlowe (1960) argument cited earlier, one would naturally predict these findings. But Webster carries his interpretation further:

> Finally . . . it becomes clearer why *RS* is a measure of inhibition; both these correlating scales measure lack of inhibition or control. In particular, *Schizoid Functioning* measures a kind of ego-diffusion which is very typical of the undercontrolled college student (p. 5).

In summary, there appears to be a general agreement in the literature that there is a trait of response acquiescence, and that it is probably closely related to some personality variable. There is also high agreement as to what to call this variable, or what kind of dimension to put it on: acquiescers are stimulus accepting, uninhibited, conformers; nonacquiescers are stimulus rejecting, inhibited, independents.

The Present State of Affairs

Throughout their provocative discussion of acquiescence and personality variables, Couch and Keniston moved progressively further up the abstraction ladder. Beginning with individual item responses they moved via factor analysis, projective testing, and depth interviewing to the level of ego functioning and "psychological inertia." Their progression was not unique, for it paralleled the route taken by Bass from the endorsement of proverbs via correlational procedures to "Babbittism!" This comment is not intended to take issue with the language used by former investigators, nor to debunk the design of their research. Such language, while abstract, nevertheless communicates ideas and feelings to a large audience of psychologists, particularly clinicians. Likewise such research is an integral part of the *inductive* process of theory building, which is a valuable means to an end.

However, left in this present state, the task is unfinished. To assume that the personality correlates of response acquiescence have been identified is to make the present collection of inductive research findings and end in itself. The task remaining should be obvious, i.e., the *deductive* formulation and testing of hypotheses to predict the behavior which the theory indicates should be related to the stylistic variables.

This review of the literature has been organized around the author's impression that what has gone on in the past has resulted primarily in descriptive information. The question proposed, and answered with progressively more rigor and precision, appears to have been: "What is the acquiescent person like?" There may be those who would argue that this is an inappropriate question to ask. It is only a minor variation of the "What is . . . ?" type of question which Muenzinger (1957) labeled as a "sterile exercise" in psychological research. How this argument is to be resolved will be left for the reader to decide for himself. The point to be made here is based on the assumption that the previously gathered descriptive data do represent a valuable contribution to the area of personality assessment. But, it is now time to shift into low gear and change course so as to proceed down the abstraction ladder in the direction of observable behavior. A question of major importance for future research is one stated in predictive form: "What will the acquiescent person do?"

At this point a short definition of terms is necessary to establish the proper *Einstellung* for the point to be made in the following section. The

problem of definition revolves around the usage of the term "observable behavior." It is essential to distinguish behavior in situations especially devised to observe a subject's responses in the laboratory, or in field settings, from patterns of responses made to paper-and-pencil questionnaires or inventories. There is no argument, certainly, that the term observable behavior, broadly conceived, includes both activities. Perhaps a useful distinction to some would be in terms of "psychometric versus nonpsychometric" situations. Yet, again, broadly defined, any measurement of behavior is a psychometric situation. There is a meaningful distinction between the behavior involved in reporting the distance traveled by a stationary light in an autokinetic conformity task, and the behavior involved in marking the agree category on an IBM answer sheet. Certainly there are few who would not grant this distinction, or who would not further grant that the distinction is based on the methodological independence of the two situations. In the following paragraphs the term observable behavior is used to denote responses elicited in a laboratory task as distinct from those elicited by the standard psychometric instrument.

Suggested Considerations for Future Research

Measures of Personality Variables

The first point to be made in this regard has already been alluded to. There is remarkable dearth of studies in this area which have attempted to study the relationship between response style measures and observable behavior measures of personality variables. The investigations of Crowne and Marlowe and their students are a notable exception. In the area of acquiescence only Jackson has used an independent behavioral measure of the trait he was studying along with response style. The reasons for the needed shift to behavioral tasks are clear. Of primary importance is the fact, well recognized by most, that paper-and-pencil personality scales are heavily loaded with RS influence. To the extent that this is true, it naturally leads to inflated correlations between an RS measure, A, and a personality scale, B. If item content is considered of minor importance in determining a scale score, as RS research generally assumes, then correlating A with B to show the relationship of two response style measures is a reasonable procedure. But to correlate A with B and conclude a relationship between response style and the trait purportedly measured by the content of B is a logically unjustifiable procedure. The paper by Webster (1960) which was discussed above is a good example of this remarkably inadequate approach to hypothesis testing. Personality inventories and RS measures are, by nature of their similar construction, to say nothing of item overlap, highly contaminated methodologically. Because of the generalized operation of RS variables, one cannot serve as an independent criterion for the other. When Frederiksen and Messick (1958) corrected their personality scale scores for RS , they found quite low relationships. Independently observed behavior is the only meaningful criterion measure of the personality variables.

Finally, inasmuch as the goal of psychological research is generally accepted to be the prediction of behavior, it is inadequate to stop short of that point. Admittedly laboratory conditions which provide the necessary controls of relevant variables also produce an artificiality which makes the situation unlike the subject's real world. Yet, it is a step in

the direction of the ultimate criterion, and one which justifies the added expense and effort.

Measures of RS Variables

In the area of acquiescent responding, several instruments and techniques have been proposed. They include various kinds of content from aphorisms to statements of personal and social attitudes. Some instruments attempt to measure pure acquiescence by putting an individual in a situation where he is forced to respond to ambiguous, or essentially meaningless stimuli. Cronbach (1950) suggested that response tendencies should be most apparent in situations where stimulus conditions are most uncertain. Berg and Rapaport (1954) confirmed this expectation by showing that consistent preferences for certain response options result when the subjects respond to an "unstructured questionnaire" wherein they guess about the nonexistent items the experimenter is supposedly reading to them. Bass (1956) also used this unstructured technique and compared it with his content-laden Social Acquiescence Scale. He found a correlation of .00.

If acquiescent behavior is to be interpreted as the indiscriminant use of yes, true, and agree options, irrespective of item content, correlations should be high between content and noncontent measures of acquiescence. Or, if there are two types of acquiescence as Bass has suggested, one would at least expect that various noncontent measures would constitute one of the types, and correlate highly with one another. McGee (1962) has found correlations which suggest that this is not true either.

While this discussion has focused only on the measurement of acquiescence, the point applies equally well to any response style one wishes to relate to central personality traits. Consequently, the course of future research on the personality correlates of response styles is clearly indicated. The basic assumption is still in need of verification. But first, two things must be demonstrated: that techniques are actually available for measuring a "pure" response style tendency, and that this response style variable can be used to predict behavior in an independent situation on the basis of some theoretical interpretation of that variable. Until these two points are demonstrated there is no defensible evidence that test taking response style is related to underlying personality syndromes, or traits.

Summary

The recent surge of interest in response style components of personality tests scores has led to a more specific interest in measures of response variables as predictors of underlying personality traits of the responders. The research studies relevant to this question, most of which appeared in the literature since 1958, have been reviewed. The point was made that these investigations have provided meaningful abstract descriptions of the personalities of individuals with certain response style tendencies, but little real defensible data to tie response styles to the criterion of independently measured behavior. Suggestions were made for designing future research efforts such that the data will lead to a prediction that the acquiescent individual will *do* something in a particular situation as well as merely say "yes" more times than he says "no" on the F Scale. Only with such data is it felt that an adequate criterion will exist for claiming a relationship between response tendencies and basic personality traits.

REFERENCES

Adorno, T. S., Frenkel-Brunswik, Else, Levinson, D. J., & Sanford, R. N. *The authoritarian personality.* New York: Harper, 1950.

Allison, J., & Hunt, D. E. Social desirability and the expression of aggression under varying conditions of frustration. *J. consult. Psychol.*, 1959, **23**, 528–532.

Asch, S. E. Studies of independence and submission to group pressure: I. A minority of one against a unanimous majority. *Psychol. Monogr.*, 1956, **70**(9, Whole No. 416).

Barnes, E. H. The relationship of bias test responses to psychopathology. *J. abnorm. soc. Psychol.*, 1955, **51**, 286–290.

Barron, F. Some personality correlates of independence of judgement. *J. Pers.*, 1953, **21**, 287–297.

Bass, B. M. Authoritarianism or acquiescence? *J. abnorm. soc. Psychol.*, 1955, **51**, 616–623.

Bass, B. M. Development and evaluation of a scale for measuring social acquiescence. *J. abnorm. soc. Psychol.*, 1956, **53**, 296–299.

Bass, B. M. Reply to Messick and Jacksoh's comments on authoritarianism or acquiescence. *J. abnorm. soc. Psychol.*, 1957, **54**, 426–427.

Bendig, A. W. "Social desirability" and "anxiety" variables in the IPAT anxiety scale. *J. consult. Psychol.*, 1959, **23**, 377.

Berg, I. A. Response bias and personality: The deviation hypothesis. *J. Psychol.*, 1955, **40**, 62–72.

Berg, I. A. Deviant responses and deviant people: The formulation of the deviation hypothesis. *J. counsel. Psychol.*, 1957, **4**, 154–161.

Berg, I. A. The unimportance of item content. In B. M. Bass and I. A. Berg (Eds.), *Objective approaches to personality assessment.* New York: Van Nostrand, 1958.

Berg, I. A. Measuring deviant behavior by means of deviant response sets. Paper read at Symposium on experimental clinical psychology, University of Virginia School of Medicine, 1959.

Berg, I. A., & Collier, J. S. Personality and group differences in extreme response sets. *Educ. psychol. Measmt.*, 1953, **13**, 164–169.

Berg, I. A., Hunt, W. A., & Barnes, E. H. *The perceptual reaction test.* Evanston, Ill.: Author, 1949.

Berg, I. A., & Rapaport, G. M. Response bias in an unstructured questionnaire. *J. Psychol.*, 1954, **38**, 475–481.

Chapman, L. J., & Campbell, D. T. Response set in the F scale. *J. abnorm. soc. Psychol.*, 1957, **54**, 129–132.

Cohn, T. S. The relation of the F scale to a response set to answer positively. *Amer. Psychologist*, 1953, **8**, 335. (Abstract)

Couch, A., & Keniston, K. Yeasayers and naysayers: Agreeing response set as a personality variable. *J. abnorm. soc. Psychol.*, 1960, **60**, 151–174.

Cowen, L. C., & Tongas, P. The social desirability of trait descriptive terms: Applications to a self-concept inventory. *J. consult. Psychol.*, 1959, **23**, 361–365.

Cronbach, L. J. Response sets and test validity. *Educ. psychol. Measmt.*, 1946, **6**, 475–494.

Cronbach, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt.*, 1950, **10**, 3–31.

Crowne, D. P., & Marlowe, D. A new scale of social desirability independent of psychopathology. *J. consult. Psychol.*, 1960, **24**, 349–354.

Edwards, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *J. appl. Psychol.*, 1953, **37**, 90–93.

Festinger, L., & Carlsmith, J. M. Cognitive consequences of forced compliance. *J. abnorm. soc. Psychol.*, 1959, **58**, 203–210.

Frederiksen, N., & Messick, S. Response set as a measure of personality. Technical Report, 1958, Educational Testing Service, Princeton, New Jersey, Office of Naval Research Contract Nonr-694(00).

Gage, N. L., & Chatterjee, B. B. The psychological meaning of acquiescence set: Further evidence. *J. abnorm. soc. Psychol.*, 1960, **60**, 280–283.

Gage, N. L., Leavitt, G. S., & Stone, G. C. The psychological meaning of acquiescence set for authoritarianism. *J. abnorm. soc. Psychol.*, 1957, **55**, 98–103.

Grigg, A. E., & Thorpe, J. S. Deviant responses in college adjustment clients: A test of Berg's deviation hypothesis. *J. consult. Psychol.*, 1960, **24**, 92–94.

Hardy, E. Predicting achievement in the school of agriculture. Unpublished master's thesis, Kansas State University, 1956.

Helmstadter, G. C. Procedures for obtaining separate set and content components of a test score. *Psychometrika*, 1957, **22**, 381–394.

Jackson, D. N. Stability and resistance to field forces. Unpublished doctoral dissertation, Purdue University, 1955.

Jackson, D. N. Independence and resistence to perceptual field forces. *J. abnorm. soc. Psychol.*, 1958, **56**, 279–281.

Jackson, D. N. Cognitive energy level, response acquiescence, and authoritarianism. *J. soc. Psychol.*, 1959, **49**, 65–69.

JACKSON, D. N. Stylistic response determinants in the California psychological inventory. *Educ. psychol. Measmt.*, 1960, **20**, 339–346.

JACKSON, D. N., & MESSICK, S. J. A note on "ethnocentrism" and acquiescence response sets. *J. abnorm. soc. Psychol.*, 1957, **54**, 132–134.

JACKSON, D. N., & MESSICK, S. J. Content and style in personality assessment. *Psychol. Bull.*, 1958, **55**, 243–252.

JACKSON, D. N., & MESSICK, S. J. Acquiescence and desirability as response determinants on the MMPI. *Educ. psychol. Measmt.*, in press.

JACKSON, D. N., MESSICK, S. J., & SOLLEY, C. M. How "rigid" is the "authoritarian?" *J. abnorm. soc. Psychol.*, 1957, **54**, 137–140.

JACKSON, D. N., & PACINE, L. Response styles and academic achievement. Research Bulletin No. 7, 1960, Pennsylvania State University, Department of Psychology.

LEAVITT, H. J., HAX, H., & ROCHE, J. H. "Authoritarianism" and agreement with things authoritative. *J. Psychol.*, 1955, **40**, 215–221.

LENTZ, T. F. Acquiescence as a factor in the measurement of personality. *Psychol. Bull.*, 1938, **35**, 659. (Abstract)

MCGEE, R. K. The relationship between response styles and personality variables: I. The measurement of response acquiescence. *J. abnorm. soc. Psychol.*, 1962, **64**, 229–233.

MARLOWE, D., & CROWNE, D. P. Social desirability and response to perceived situational demands. *J. consult. Psychol.*, 1961, **25**, 109–115.

MESSICK, S. J. Dimensions of social desirability. *J. consult. Psychol.*, 1960, **24**, 279–287. (a)

MESSICK, S. J. Response style and content measures from personality inventories. (Res. Bull. No. 60-12) Princeton, N. J.: Educational Testing Service, 1960. (b)

MESSICK, S. J., & FREDERIKSEN, N. Ability, acquiescence, and "authoritarianism." *Psychol. Rep.*, 1959, **4**, 687–697.

MESSICK, S. J., & JACKSON, D. N. Authoritarianism or acquiescence in Bass' data. *J. abnorm. soc. Psychol.*, 1957, **54**, 424–426.

MESSICK, S. J., & JACKSON, D. N. Aquiescence and the factorial interpretation of the MMPI. *Psychol. Bull.*, 1961, **58**, 299–304.

MUENZINGER, K. F. Introduction. In, *Contemporary approaches to cognition: Colorado symposium on cognition.* Cambridge: Harvard Univer. Press, 1957.

SAUNDERS, D. R. Some preliminary interpretive material for the PRI. (Res. Memo. No. 55-15) Princeton, N. J.: Educational Testing Service, 1955.

SECHREST, L. B., & JACKSON, D. N. Deviant response tendencies: Their measurement and interpretation. Paper read at American Psychological Association, Chicago, September 1960.

STRICKLAND, BONNIE. A new look at social desirability. Paper read at Ohio Psychological Association, Columbus, April 1960.

TAYLOR, J. B. Social desirability and MMPI performance: The individual case. *J. consult. Psychol.*, 1959, **23**, 514–517.

TAYLOR, J. B. What do attitude scales measure: The problem of social desirability. *J. abnorm. soc. Psychol.*, 1961, **62**, 386–390.

WAHLER, H. J. Social desirability and self-ratings of intakes, patients in treatment, and controls. *J. consult. Psychol.*, 1958, **22**, 357–363.

WEBSTER, H. The meaning of "response set" in personality inventories. Paper read at American Psychological Association, Chicago, September 1960.

(Received January 30, 1961)

Psychological Bulletin
1962, Vol. 59, No. 4, 296-300

A NOTE ON THE INCONSISTENCY INHERENT IN THE NECESSITY TO PERFORM MULTIPLE COMPARISONS

WARNER WILSON
University of Hawaii

Some studies involve only two groups and provide only one difference to be tested for significance. Other studies involve several groups and provide many differences to be tested for significance. A question has arisen in the literature (Duncan, 1955; Ryan, 1959; Tukey, 1949) as to how significance should be determined when a number of tests are to be made in the same experiment. Ryan (1959) has performed a valuable function by pointing out that there are several ways of dealing with this problem.

It would be possible to adopt a strategy that would hold errors constant per comparison, per hypothesis, per experiment, per group, or even per subject. The question is essentially: what is the appropriate unit in which to evaluate research? It is the thesis of this paper that the most defensible decision is to divide our work into separate tests of hypotheses and to hold constant the expected number of errors per hypothesis tested.

The number of groups involved in the test of a single hypothesis may vary depending on the attitude of the experimenter and the nature of the hypothesis. Often an experiment determines the effects of several degrees of a measurable variable: in this case the hypothesis is usually that there is some relationship between an independent and dependent variable. In this case differences between individual groups may be of little concern. For example, if length of food deprivation is varied at 2-hour intervals from 2 to 24 hours, it is the overall variability between groups that is of interest, not the difference between any particular pair of groups. A failure to find a difference between the 8-hour and 10-hour group would be of little importance. In other cases several groups may be run that do not represent points on a measurable dimension and in such cases the difference between each group and every other group may be viewed as a separate hypothesis. For example, if the results of five different therapies are compared, the significance or nonsignificance of the difference between any two groups would probably be considered important. In this second case there would be more hypotheses but less data relevant to each one. If several variables are studied in a single experiment the significance of the effect of each variable and each interaction may be tested as a separate hypothesis. The practice of holding errors constant per hypothesis tested seems to be by far the most common in the literature: the *F* test is typically employed when the performance of several groups is subsumed under one hypothesis, and the *t* test is typically used to test differences between pairs of groups when each pair is construed as bearing on a separate hypothesis. Many, if not most, researchers are not even aware of the various special statistics that have been devised for the purpose of using some unit other than the hypothesis as the basis for error rate.

It is necessary to recognize, however, that all discussions in the

literature recommend some unit other than the hypothesis as the basis for determining error rates. Ryan (1959) and Tukey (1953 unpublished), for example, favor the experiment as the preferred unit. The only dissenter to this general approach seems to be Duncan (1955) who favors what is essentially a compromise position. The purpose of this paper is to consider the pros and cons of the per-experiment versus the per-hypothesis approach. An attempt is made to make clear that some inconsistency is involved in either case and that a consequence of this fact is that several of the arguments offered in favor of the per-experiment strategy are in fact offset by parallel, equally logical arguments, in favor of the per-hypothesis strategy. It is pointed out below that while it is impossible to prefer one approach to the other on logical grounds, other considerations actually favor the per-hypothesis approach. Ryan (1959) and Tukey (1953 unpublished) actually speak of a per-comparison (rather than a per-hypothesis) approach as the possible alternative to the per-experiment approach. Although the two may seem to be similar, the per-hypothesis approach is different from the per-comparison in that any number of comparisons may be considered in testing one hypothesis, however, the arguments presented in relation to the per-comparison strategy apply in exactly the same way to the per-hypothesis approach.

As Ryan makes clear, if a per-hypothesis strategy is used, the same number of errors will be expected in 100 small experiments, each of which tests one hypothesis, as will be expected in a large experiment that tests 100 hypotheses (Ryan, 1959, pp. 30–34). Ryan maintains that independence of the tests or lack of it makes no difference: "The error rates per comparison and per experiment are completely unaffected by independence or lack of it" (Ryan, 1959, p. 34). Obviously if the error rate per hypothesis is held constant, the error rate per experiment will vary, depending on the size of the experiment. On the other hand, if the error rate per experiment is held constant, the error rate per comparison will vary, depending again on the size of the experiment. Since inconsistency is involved in either case a choice on purely logical grounds does not seem possible. If the implications of this fact are followed consistently, several of the arguments in favor of the per-experiment solution become meaningless.

Ryan (1959) and Tukey (1953 unpublished) both argue that a per-hypothesis strategy implicitly gives a person license to make relatively more errors per experiment merely because he has been industrious in running many groups. Although this argument seems quite irrelevant to the issue, it is only fair to note the other side of the question. The per-experiment strategy implicitly gives a person license to make relatively more errors per hypothesis, merely because he has been lazy, as evidenced by the running of few groups! It is hard to see how the first argument can be considered more compelling than the second.

Ryan (1959) also argues that a per-hypothesis strategy, by favoring the person who is industrious, as evidenced by the running of many groups, may lead people who run many subjects in a two-group experiment to demand the privilege of using a higher error rate because they too have been industrious, as evidenced by the running of many subjects. While this argument seems a little too artificial to deserve con-

sideration, it is once again easy to point out the parallel counter-argument. The per-experiment solution, by favoring the person who is lazy, as evidenced by the running of few groups, might lead those who run few subjects in a two-group experiment to demand the privilege of using a higher error rate because they too have been lazy! Once again it is hard to argue that the possible consequences of the per-hypothesis approach are worse than the possible consequences of the per-experiment approach.

Some of the comments in the literature (e.g., Ryan, 1959, pp. 35–37) may suggest that the use of a per-hypothesis strategy necessarily results in an inordinate amount of error or at least in more errors than a per-experiment strategy. Such a conclusion would be completely false. It is true that if a per-hypothesis error rate is employed there will be *relatively* more errors *per experiment* in *large* experiments, but it is also true that if a per-experiment error rate is employed there will be *relatively* more errors *per hypothesis* in *small* experiments. The total expected number of errors can be controlled equally well no matter in what unit results of research are measured. Insistence on fewer errors per-experiment would decrease total errors to be sure, but insistence on fewer errors per-hypothesis would decrease total errors equally well. Ryan actually concedes this point at one place, but apparently fails to recognize its implications (Ryan, 1959, pp. 37–38). Unless one wishes to argue that an error does more damage merely because it occurs in a large experiment, it must be concluded once more that there is no logical basis on which to choose between the different strategies.

The writer firmly agrees with those who think a more rigorous control of errors is called for; however, he suggests that the most effective way for workers to achieve this is to hold the expected error rate constant at .001 per hypothesis. Suppose one person publishes at the .05 level per experiment and a second publishes at the .001 level per hypothesis. Assuming that the second person's experiments test less than 50 hypotheses on the average, he will make fewer errors both per experiment and per hypothesis than will the first person. Clearly an experimenter can be as rigorous as he wishes and still use the hypothesis as his research unit.

Another type of consideration relates to the effect that each strategy might have on the behavior of researchers as they design, carry out, and write up experiments. It has been argued (Ryan, 1959, p. 36) that a per-hypothesis type approach encourages investigators to include "irrelevant" variables in their studies merely to increase their chances of obtaining one or more "significant" findings to publish. Surely such motivation is deplorable. However, it is doubtful that many researchers will deliberately resort to such tactics, and surely editors will be reluctant to accept implausible false positives no matter what statistical techniques are used. Furthermore the line between adding irrelevant variables and exploring new possibilities is rather subtle and it is not at all certain that psychology would not profit from some additional blind seeking for relationships. It is necessary to insist on looking at both sides of the picture. What sort of pressures does the per-experiment procedure apply to the researcher? It seems likely that, for better or worse, most experimenters design studies to demonstrate relationships they believe to exist. Their desire is

to obtain data that will support their hypotheses and compel others to accept them. Very generally it can be assumed that there is often a choice between testing a number of hypotheses in different experiments by running only the two groups expected to be most extreme versus testing several hypotheses in one experiment by running several groups to determine the effects of each variable.

The latter, more extensive type of study, is greatly to be preferred since it consumes less journal space per hypothesis, it allows for the evaluation of interaction effects, and it gives some idea of the shape of relationships. The per-experiment approach seem to discourage extensive studies because the more extensive the study the less the likelihood of being able to accept any given hypothesis as correct. In other words if a per-experiment strategy is used, the smaller the pieces in which one can publish, the greater his chances of having significant findings to report. When a per-hypothesis strategy is followed this additional encouragement to publish in small pieces is not present. The literature is currently cluttered with small one-shot studies and there is a relative dearth of well conceived, intensive investigations. Certainly all angles should be considered before a strategy is advocated that might intensify this unfortunate tendency. Apparently either the per-experiment or the per-hypothesis strategy might have ill effects on certain researchers, but once again it is hard to see the arguments in favor of the per-experiment approach as more compelling than those favoring the per-hypothesis approach.

In addition it can be pointed out that there are strong advantages to the per-hypothesis solution. The basic question is, what is the most meaningful unit in which to evaluate research? Traditional practice apparently has chosen the hypothesis as the unit and this paper maintains that this is the correct choice. It seems that the hypothesis is psychologically the more logical unit. This writer, at least, would prefer to be confronted with a great array of findings, all of which (statistically speaking) have a comparable probability of being correct, rather than to be confronted with a number of conclusions each of which can be accepted with more or less confidence depending on the size of the experiment from which they were derived.

Another major advantage of the per-hypothesis approach is the fact that it requires no additional learning on the part of researchers. Obviously the more complicated statistics become the more time it will take to learn to use them and the less time will be available for research itself. It seems foolish for researchers to accept additional statistical complications unless there are telling reasons for doing so. It might also be added that it is practically impossible for a statistically naive researcher to abandon the traditional per-hypothesis techniques because statisticians have not yet agreed upon any other strategy or even on how best to achieve the various alternatives that have been advocated. Duncan (1955) mentions nine different solutions to the problem of multiple comparisons and comments that, "Unfortunately, these tests vary considerably and it is difficult for the user to decide which one to choose for any given problem" (p. 2). One purpose of Duncan's article was to propose still another solution: It has not received general acceptance (Ryan, 1959) and it seems apparent that statisticians have no generally agreed upon alternative to suggest as

a possible replacement for the per-hypothesis approach.

It must be concluded that the arguments in favor of the per-hypothesis strategy are more numerous and more compelling than those in favor of the per-experiment solution. Therefore the less effortful per-hypothesis approach should be continued indefinitely unless valid arguments are presented in favor of a different strategy.

REFERENCES

DUNCAN, D. B. Multiple range and multiple *F* tests. *Biometrics*, 1955, **11**, 1–42.

RYAN, T. A. Multiple comparisons in psychological research. *Psychol. Bull.*, 1959, **56**, 26–47.

TUKEY, J. W. Comparing individual means in the analysis of variance. *Biometrics*, 1949, **5**, 99–114.

(Received February 28, 1961)

Psychological Bulletin
1962, Bol. 59, No. 4, 301–305

THE EXPERIMENT AS THE UNIT FOR COMPUTING RATES OF ERROR

THOMAS A. RYAN
Cornell University

I am very glad that Wilson (1962) has addressed himself to the basic issues involved in multiple comparisons—issues partly of logic and partly of research strategy. There is a very real dilemma involved, and one which needs to be brought into the open even if we cannot reach a single solution. In my discussion of the problem (Ryan, 1959b), I believed that the balance of the arguments favored the error rates based upon the experiment as the unit, and I stated that conclusion in its strongest form. Perhaps the statement was one-sided, as Wilson believes, but I hoped that this would bring the issue more clearly to the readers than a less positive statement. Wilson has chosen the other horn of the dilemma and has done a service in stating the case for his choice so clearly. Many of my colleagues had previously expressed unhappiness with my conclusions, not on logical grounds, but because experiment-based error rates made life more difficult for the researcher, who must find bigger *t*'s for significance.

Even a casual examination of the current journals will show that this is an issue which crops up in a very high percentage of research reports, though usually unrecognized by the authors of the reports. Wilson's (1962) arguments (apart from his plea for more stringent significance levels) tend to support the status quo, but it would be a pity if his position were adopted simply out of inertia without careful weighing of the advantages and disadvantages. A thorough examination of the issues might even lead us to abandon significance testing in favor of some more useful form of statistical treatment, although I do not yet see what might replace significance tests or confidence limits.

The issues must be met and settled by psychologists themselves, or by other "consumers" of statistical method. The statisticians can tell us how to accomplish what we want to do, but we must decide what we want to do in terms of overall research strategy. If we could quantify the costs of doing research with various experimental designs, the "earnings" due to correct conclusions and the "losses" due to Type I and Type II errors, the whole problem could be solved mathematically. Since, clearly, this can be done only in such limited and artificial situations that it could not provide a general procedure, we are forced to choose our procedures on the basis of broad and qualitative arguments.[1] This is what we do when we choose a significance level for a single comparison, since we are balancing qualitatively the risks of Type I error against the risks of Type II error. The same kind of qualitative balancing is necessary when we consider the issue of error rates per hypothesis versus error rates per experiment. Unfortunately, however, the balancing of risks becomes much more complex than it is

[1] Tukey (1960) has recently made a distinction between "decisions" and "conclusions," the latter being of more relevance to scientific work. He also argues that the theory of statistical decision is not appropriate to the testing of conclusions.

for the simple, isolated comparison.

I have asked for the opportunity to comment on Wilson's (1962) paper, not because I want, or expect, to *prove* that he is wrong, but because some of the implications of his argument need to be pursued further. His conclusion may be the best one, but I am not yet convinced.

Let us first stipulate (*a*) an important point of agreement and (*b*) one issue which can profitably be left for a separate discussion:

I would agree that there are many cases where overall analysis of variance is more appropriate than multiple comparisons of individual means. These are cases which are essentially problems of regression, which I intentionally left out of my earlier analysis of the problem of multiple comparisons (see Ryan, 1959a, p. 396). Even here the issue of error rates may rear its head if we wish to test separately for linear, quadratic, and higher order components, or if we wish to state where the maximum or minimum falls. This latter problem becomes closely similar to the problem of multiple comparisons, so I shall not argue it separately.

I shall leave aside the question of whether the various F tests in a complex analysis of variance should be treated with an error rate per F test (per hypothesis), or on the basis of an overall rate for the whole experiment. Here I only wish to make clear that the point of view I previously expressed on this problem (Ryan, 1959b, p. 44) is not to be attributed to Tukey. Tukey prefers to follow current standard practice in complex analysis of variance, allocating an error rate to each F separately. If one of the variates is subjected to multiple comparisons, he would allow this *family* of comparisons the same error rate *familywise* that would otherwise have been allocated to the single F test for this variate. When I stated that the same arguments which lead to a familywise control of error could be used to support an overall control of error for the whole multivariate experiment, this was my own conclusion, which Tukey did not accept.[2] For the present let us leave the multivariate problem out of the discussion as too complex to deal with until we have settled the more basic question of how to deal with the univariate experiment.

Turning now to what appears to be the principal point of difference, Wilson objects to the argument that the control of errors per hypothesis gives the experimenter more chance of finding some significant differences merely by being more diligent and studying more different conditions. I will admit that this is a question of values (but not of morals). There are several questions of values involved throughout significance testing; e.g., choosing to work at the .01 level instead of the .05 level is also a question of value, or how important we consider erroneous conclusions to be. Wilson is justified in questioning my point because the problem was incompletely analyzed in my earlier discussion. At the time I was merely trying to express the rather vague notion that obtaining significant results should not depend *solely* upon the persistence or diligence of the experimenter. These are admirable qualities, but a statement of significance should also bear some relation to the facts of nature, as well as to the diligence of the experimenter in seeking out these facts.

It must be emphasized that error rates refer to what happens when the null hypothesis is true. If the experimenter is so perspicacious or so lucky

[2] J. W. Tukey, personal communication, 1956.

as to find real effects upon behavior, the Type I error rates no longer apply and considerations of power enter the picture. If, on the other hand, he is so unfortunate as to waste his energy upon a real null situation, there is no reason to allow him to make some mistakes as a consolation prize.

I will agree with Wilson, however, that this whole aspect of the relation of error rates to experimental investment needs more exact analysis. Unfortunately there are so many facets to be considered at once that we have to oversimplify to make any sense of the problem. One approach to the problem is to try to hold the factor of experimental effort or cost constant while comparing different experimental designs. One approximation would be to consider the total number of observations as a measure of the work done in the experiment. This might be a reasonable assumption on the average, since we are concerned with the choice of designs to be used with the same kind of measures. Suppose for example, that Experimenter A studies two conditions with 100 observations per condition. Experimenter B also makes 200 observations altogether but spreads them over 10 different conditions of the variable, and compares each mean with each of the others. A tests 1 null hypothesis, while B tests 45 hypotheses with the same total amount of data. If the error rate is controlled per hypothesis, B can be expected to make .45 errors while A is making .01 errors, but their error rates will be equalized if the experiment is used as the base for computing error rate. Thus it seems that the rate of error *per experiment* should be controlled if we wish to equalize the risks of Type I error for a given amount of experimental data spread over varying numbers of groups.

Now consider Experimenter C who also studies 10 conditions, but collects a greater amount of data so that the additional groups represent additional experimental effort. He is, as Wilson points out, not allowed any more error on the per experiment basis. *But this is also true if error rate is computed per hypothesis.* Specifically, suppose that B makes 200 observations spread over 10 conditions, while C makes 1,000 observations on the same 10 conditions. Both methods of computing error rate would treat the two experimenters alike. One would allow both to make .45 errors the other would allow both to make .01 errors.

C is not allowed any more Type I errors than B for his extra effort by any of the methods of computing rate of error, but C does gain in power from the extra observations. This is consistent with current practice, in that the error rate for A's single comparison would not be changed if he collected more data for the two conditions. In short, controlling errors per experiment holds the amount of error constant for a fixed amount of experimental effort whether it is devoted to a single pair of conditions or many different conditions. Controlling the rate of error per hypothesis allows the error rate to increase as the number of groups increases, even if the same total amount of experimental effort is spread thinly over many groups. For both methods, additional observations, without a change in the research design, are used to increase the power of the experiment but do not change the rate of Type I error.

The above argument points up the fact that there is an arbitrary decision involved in current practice even with single comparisons. It has been decided that power shall vary with number of observations, but that rate of Type I error shall not. This

presumably is the historical result of the concept of significance developing before the concept of power. Actually, of course, the concept of power derives from decision theory in which the rates of Type I and Type II error would both be variable and adjusted in terms of the costs of the two types of error. Instead, however, the concept of power has been tacked on as an adjunct to the significance test, but not controlled directly because of our ignorance of the consequences of error. This is one of the aspects of present practice in significance testing which needs more thorough examination. For example, do we really want extra effort in research to be devoted to the detection of smaller and smaller differences? Meanwhile, if we accept current practice as the appropriate approach to single comparisons, the comparable solution to multiple comparisons would be to control the error rate per experiment.

There is another argument in favor of basing error rate upon the single hypothesis, an argument which Wilson does not mention directly but is related to his fear of discouraging large-scale experiments if we adopt the error rate per experiment. This is the point that even the experimenter who tests a single hypothesis in an experiment is one of many who may be working upon the problem over the years, or he may be doing one experiment in a series of his own. Yet (the argument continues) it is accepted practice for him to test this single hypothesis as though it were isolated from all of the studies carried out by him or others in the past. If he is allowed to do this, why should the experimenter who tests several hypotheses in the same paper be penalized by requiring him to limit his errors for the total experiment rather than for each hypothesis separately?

This is a very powerful argument if we accept current practice as appropriate, and if we agree that current practice is as described. I would question both of these assumptions, however. An experimenter never considers his results in isolation from the rest of the available data in the field. One result which is out of line with other findings in the field is likely to be regarded with suspicion even if it is technically significant, and further replication will probably be called for. Even though this is not a quantified or explicit procedure it is in the same spirit as controlling errors for the total experiment in multiple comparisons. We are handicapped in our knowledge of the total experimental background because of the failure to publish many negative findings and the consequent bias in the results available to us (see Sterling, 1959). Nevertheless we do, and should, try to take account of total mass of information available to us in interpreting any specific experimental result.

Wilson's (1962) fears that experimenters will be discouraged from doing large experiments with many different conditions if we expect them to limit the total errors for the whole experiment. Yet why should they be discouraged? They are doing these experiments because they want to find real effects, not because they want to report Type I errors. Wilson seems to believe that I advocated error rates per experiment because I wanted to increase the stringency of our standards of significance. Advocating the experiment as the proper base for computing error rates does not imply that we should set up more stringent criteria of significance than are now customary; this is a

separate and independent question. It is true that a .01 level of significance experimentwise or per experiment means a much lower probability level for individual comparisons within the experiment if many comparisons are made. We do not have to work at the .01 level experimentwise, however. The problem is not what particular probability should be chosen, but which method of computing the rate is comparable from one research to another. I have merely argued that the rates per experiment or experimentwise, whether .01 or .90, provide greater comparability from one research to another.

It happens that I, like Wilson, do believe that we should use more stringent criteria of significance if we use significance tests at all. But he is quite correct in pointing out that greater stringency could be achieved by lowering the probability levels for errors per hypothesis as well as controlling errors per experiment. Consequently, the choice of error rates is logically irrelevant to the issue of stringency. My own reasons for supporting greater stringency are based upon the belief that Type I errors are more dangerous in the present state of development of psychology than are Type II errors. In other words, I believe that it is less important if we miss some very small effect of a variable, than it is to claim that the variable has an effect (of unspecified magnitude) which does not actually exist at all. This is however, another problem of many facets which cannot be threshed out here.[3]

To summarize, Wilson (1962) has presented some very strong arguments for controlling error rates per hypothesis instead of for the whole experiment. It is a service to have this side of the issue presented so clearly, and it is possible that he is right. There are, however, strong counterarguments which I have tried to present, and which still weigh heavily enough to convince me that we should control the error rates per experiment. To me, the strongest argument is that controlling the rate of error per hypothesis permits wide variation in the total amount of error expected for different experimental designs which involve the same total number of observations.

The issue is by no means settled, however. There are many factors which must be weighed against each other, and there are probably some considerations that have not yet been dealt with. An adequate solution of the problem might even lead to an abandonment of significance testing in favor of some other method of dealing with the effects of sampling error which would not create the dilemma with which we are now faced.

[3] One especially important problem is what we shall do with negative results (see Sterling, 1959; Tullock, 1959).

REFERENCES

RYAN, T, A. Comments on orthogonal components. *Psychol. Bull.*, 1959, **56**, 394–396. (a)

RYAN, T. A. Multiple comparisons in psychological research. *Psychol. Bull.*, 1959, **56**, 26–47. (b)

STERLING, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Amer. Statist. Ass.*, 1959, **54**, 30–34.

TUKEY, J. W. Conclusions vs. decisions. *Technometrics*, 1960, **2**, 423–433.

TULLOCK, G. Publication decisions and tests of significance: A comment. *J. Amer. Statist. Ass.*, 1959, **54**, 593.

WILSON, W. A note on the inconsistency inherent in the necessity to perform multiple comparisons. *Psychol. Bull.*, 1962, **59**, 296–300.

(Received March 17, 1961)

Psychological Bulletin
1962, Vol. 59, No. 4, 306–310

AN EXACT MULTINOMIAL ONE-SAMPLE TEST OF SIGNIFICANCE

ALPHONSE CHAPANIS
Johns Hopkins University[1]

Many psychological studies yield nominal data—numbers of subjects, objects, or responses—distributed into two or more mutually-exclusive categories. With data of this type the experimenter, and the reader, usually want to know if the observed frequencies differ significantly from what one would expect on the basis of chance. Chi square can be used for making such a test provided that the numbers of observed frequencies exceed certain minimum requirements. The binomial test provides an exact test of significance for samples of any size but it can only be applied to data distributed into two categories. This article shows how the multinomial distribution can be modified and used as an exact test of significance for samples of any size and for data distributed into any number of categories. Although the test described here follows closely those given by Smith and Duncan (1945, pp. 308–326) and Tate and Clelland (1957, pp. 35–36), it differs from both of them in certain important respects.

Description of the Method

The probability that a sample of data will yield the frequencies n_1, n_2, $n_3 \cdots n_k$ distributed into k categories is given by the multinomial distribution:

[1] This paper was prepared while the author was on leave (1959–60) with the Office of Naval Research in London, England.

$$\phi = \frac{(n_1 + n_2 + n_3 + \cdots + n_k)!}{n_1!\, n_2!\, n_3! \cdots n_k!} p_1^{n_1} p_2^{n_2} p_3^{n_3} \cdots p_k^{n_k} \qquad [1]$$

where the p's are the proportions with which the characters $1,2,3, \cdots k$ occur in the population.

When we use the multinomial distribution as a test of significance we assume that the null hypothesis holds, that is, that the p's are equal and that each of them is equal to $1/k$, where k is the number of categories. If, in addition, we let $N = n_1 + n_2 + n_3 + \cdots + n_k$ then Equation 1 becomes:

$$\phi = \frac{N!}{n_1!\, n_2!\, n_3! \cdots n_k!}\left(\frac{1}{k}\right)^N \qquad [2]$$

One further addition is still necessary. If the null hypothesis holds then we should make no distinction between different permutations of any outcome. To take a particular example this means that all of the following six outcomes are equivalent:

4 *A*'s, 3 *B*'s, 1 *C*
4 *A*'s, 1 *B*, 3 *C*'s
3 *A*'s, 4 *B*'s, 1 *C*
3 *A*'s, 1 *B*, 4 *C*'s
1 *A*, 3 *B*'s, 4 *C*'s
1 *A*, 4 *B*'s, 3 *C*'s

Adding this refinement to Equation 2 gives us:

$$P = \frac{k! \prod_{i=1}^{i=k} i!}{[(t_2+1)2!] \cdots [(t_i+1)i!] \cdots [(t_k+1)k!]} \frac{N!}{n_1! n_2! \cdots n_j! \cdots n_k!}\left(\frac{1}{k}\right)^N \qquad [3]$$

where:

- P = the probability of obtaining any permutation of n_1, n_2, $n_3 \cdots n_k$ frequencies
- N = the total number of observations (individuals, objects, or responses)
- k = the total number of categories into which the N observations are distributed
- i = any integer $\leq k$
- t_i = number of ties of size i among the k frequencies
- j = any one category
- n_j = the number of observations in the jth category

Note that Equation 3 without the $(\frac{1}{k})^N$ term tells us the *number of ways* in which any particular outcome (such as 4,3,1) can occur. Note also that for $k=2$, Equation 3 reduces to the form of the binomial distribution which would be used for a test of this kind.

Equation 3 merely gives us the probability that N observations will, by chance, be distributed into k categories with any particular set of frequencies n_1, n_2, $n_3 \cdots n_k$. To use Equation 3 as a test of significance we need to add to this the probabilities of all those outcomes which are even more deviant than the one observed.

Illustration of the Method

To illustrate the application of this formula I shall use some data collected by Deininger (1960). In one part of his experiment Deininger had subjects use keysets in which the keys had maximum displacements of $\frac{1}{32}$, $\frac{1}{8}$, and $\frac{3}{16}$ inch. At the conclusion of an unspecified number of trials, 12 subjects voted for the keyset they liked least: eight disliked the $\frac{1}{32}$-inch, one the $\frac{1}{8}$-inch, and three the $\frac{3}{16}$-inch. The author concludes that "the smallest displacement appears controversial, the largest unpopular and the middle the most desirable." The question we want to answer is: If the subjects really had no particular dislikes and were simply voting randomly how likely is it that we could have obtained an outcome as deviant as 8,3,1 by chance?

To apply Equation 3 note that $N=12$, $n_1=8$, $n_2=3$, $n_3=1$, $k=3$, $t_2=0$, and $t_3=0$. Inserting these identities into Formula 3 gives us:

$$P = \frac{3!1!2!3!}{[(0+1)2!][(0+1)3!]} \frac{12!}{8!3!1!}\left(\frac{1}{3}\right)^{12}$$

$$= (11{,}880)\left(\frac{1}{3}\right)^{12}$$

The simplest way to illustrate the full computation of the probability we want is to list all possible outcomes and the number of ways in which each outcome can occur (since the term $(\frac{1}{3})^{12}$ appears as a constant we can disregard it for the time being). Table 1 gives these data.

Note that we have a check on these computations since the total for Table 1 is equal to 3^{12}.

Before continuing we need to consider what we mean by outcomes "even more deviant than" 8,3,1. What we mean are all those outcomes which have an even smaller probability of occurring. Table 2 lists these in order. The total for Table 2 is 37,431 and this value multiplied by $(\frac{1}{3})^{12}$, or divided by 531,441, gives us a probability of 0.070.

To summarize, if the null hypothesis is correct we could expect an outcome as deviant as 8,3,1 to occur about 7 times in 100. According to the usual conventions we would therefore conclude that this outcome is not statistically significant.

The Case of Ties

The example given above is convenient because it is small enough for us to see all the essential computa-

tions compactly. It does not, however, illustrate one nuance of Equation 3, namely, what happens when some categories have tied observations. In another part of his experiment Deininger (1960) had 15 subjects use five different keysets—call them A, B, C, D, and E—which differed in several ways. At the conclusion of an unspecified number of trials each subject voted for the keyset he liked best. The results were 2 votes for A, 1 for B, 4 for C, 6 for D, and 2 for E. Now we want to test the significance of the outcome 2,1,4,6,2. The novel feature of these data is the two 2s.

In this case $N=15$, $n_1=2$, $n_2=1$, $n_3=4$, $n_4=6$, $n_5=2$, $k=5$, $t_2=1$ (for n_1 and n_5), $t_3=0$, $t_4=0$, and $t_6=0$. Inserting these values into Equation 3 gives us:

$$P = \frac{5!1!2!3!4!5!}{([(1+1)2!][(0+1)3!][(0+1)4!][(0+1)5!]}\,\frac{15!}{2!1!4!6!2!}\left(\frac{1}{5}\right)^{15}$$

$$= \frac{5!1!2!3!4!5!}{2!2!3!4!5!}\,\frac{15!}{2!1!4!6!2!}\left(\frac{1}{5}\right)^{15}$$

$$= \left(\frac{5!}{2!}\right)\frac{15!}{2!4!6!2!}\left(\frac{1}{5}\right)^{15} = 0.037.$$

TABLE 1

ALL POSSIBLE OUTCOMES WHEN 12 INDIVIDUALS ARE DISTRIBUTED INTO 3 CATEGORIES AND THE NUMBER OF WAYS EACH OUTCOME CAN OCCUR

Outcome	Number of ways the outcome can occur
12, 0, 0	3
11, 1, 0	72
10, 2, 0	396
10, 1, 1	396
9, 3, 0	1,320
9, 2, 1	3,960
8, 4, 0	2,970
8, 3, 1	11,880
8, 2, 2	8,910
7, 5, 0	4,752
7, 4, 1	23,760
7, 3, 2	47,520
6, 6, 0	2,772
6, 5, 1	33,264
6, 4, 2	83,160
6, 3, 3	55,440
5, 5, 2	49,896
5, 4, 3	166,320
4, 4, 4	34,650
Total	531,441

The p of 0.037 is, of course, merely the probability of getting exactly an outcome of 6,4,2,2,1 or some permutation of this outcome. The probability of an outcome *as deviant as* 6,4,2,2,1 (computed in a manner analogous to that shown in Table 2)

TABLE 2

OUTCOMES IN ORDER OF INCREASING LIKELIHOOD UP TO AND INCLUDING THE OUTCOME 8,3,1

Outcome	Number of ways the outcome can occur
12, 0, 0	3
11, 1, 0	72
10, 2, 0	396
10, 1, 1	396
9, 3, 0	1,320
6, 6, 0	2,772
8, 4, 0	2,970
9, 2, 1	3,960
7, 5, 0	4,752
8, 2, 2	8,910
8, 3, 1	11,880
Total	37,431

is 0.49. It has, in short, no statistical significance whatsoever.

A Comparison of the Exact Multinomial Test with Chi Square

As noted above, when $k=2$, the exact multinomial test given by Equation 3 reduces to the formula which would be used for an exact binomial test. It is of interest, however, to compare the outcomes of the exact test given in this article with those of the chi square approximation commonly used for this purpose. Table 3 shows all possible outcomes when 12 individuals are distributed into three categories (from Table 1) and the exact probabilities of obtaining outcomes as least as deviant as those listed. For comparison, the third column of Table 3 shows the corresponding probabilities[2] computed by the chi square formula. One reason why chi square probabilities often do not agree with those calculated by exact tests is that chi square uses a continuous distribution to approximate discrete ones. To compensate for the errors involved in this approximation statisticians often recommend applying a correction for continuity. Although corrections for continuity tend to overcompensate a little they do usually bring chi square probabilities into closer agreement with their true values. In the fourth column of Table 3 the chi square probabilities have been corrected for continuity according to the method recommended by Cochran (1952).

TABLE 3

Exact and Chi Square Probabilities of Every Possible Outcome When 12 Individuals Are Distributed into 3 Categories

Outcome	Exact probability	Chi square probability	
		Uncorrected	Corrected for continuity
12, 0, 0	0.000006	0.00001	0.00003
11, 1, 0	0.000141	0.00010	0.00030
10, 2, 0	0.00163	0.00091	0.00104
10, 1, 1	0.00163	0.00117	0.00248
9, 3, 0	0.00412	0.00525	0.00674
6, 6, 0	0.00933	0.0498	0.0725
8, 4, 0	0.0149	0.0183	0.0267
9, 2, 1	0.0224	0.00866	0.0126
7, 5, 0	0.0313	0.0388	0.0440
8, 2, 2	0.0481	0.0498	0.0725
8, 3, 1	0.0704	0.0388	0.0440
7, 4, 1	0.115	0.106	0.135
6, 5, 1	0.178	0.174	0.253
4, 4, 4	0.243	1.000	1.000
7, 3, 2	0.332	0.174	0.253
5, 5, 2	0.426	0.472	0.606
6, 3, 3	0.531	0.472	0.606
6, 4, 2	0.687	0.368	0.417
5, 4, 3	1.000	0.779	0.883

Table 3 shows some striking discrepancies between the chi square probabilities, both uncorrected and corrected, and those resulting from the exact test. Note especially the number of discrepancies in the critical areas around the 1 and 5% points. The outcome 6,6,0, for example, is significant at the 1% level by the exact test, scarcely significant at the 5% level by the uncorrected chi square test, and not significant at the 5% level by the corrected chi square test. Similar large discrepancies occur for the outcome 9,2,1; 8,2,2; and 8,3,1.

Smith and Duncan (1945) assumed, as the chi square test does, that the probability of any occurrence is proportional to the evenness of the distribution of the N observations in the k categories. For this reason they stated that zones of re-

[2] These values were obtained, by linear interpolation when necessary, from the Pearson-Hartley (1956) tables.

jection, and so the statistical significance of any outcome, would be proportional to D^2, where, in my notation,

$$D^2 = \sum\left(\frac{n_k}{N} - \frac{1}{k}\right)^2 \qquad [4]$$

For the very small example they gave ($N=5$, $k=3$) this happened to be true, but it is certainly not true in general. The outcomes 6,6,0 and 8,2,2 (7,5,0 and 8,3,1; 6,5,1 and 7,3,2; and 5,5,2 and 6,3,3) have identical values of D^2 but markedly different probabilities of occurrence by the multinomial distribution. The symmetry implicit in D^2 is, in fact, one of the reasons why the chi square test yields probabilities which differ so markedly from the true ones (Table 3).

However, the real source of the discrepancies between the two kinds of tests lies even deeper than this. Although the formula for chi square is derived from that of the multinomial distribution (for example, Kendall, 1947), at three separate points the derivation makes use of approximations which are valid only for large Ns. Table 3 shows that the cumulative effect of the errors in these approximations may be considerable when chi square is applied to data with small n's.

A Disadvantage of the Exact Test

Perhaps the chief disadvantage of the exact test described here is that it is laborious to calculate and very quickly becomes prohibitively difficult to apply when N or k become large. The first example given in this paper is relatively straightforward and not too tedious. The second example (with 15 individuals distributed into five categories), however, required the computation of 84 separate outcomes with a total of 30,517,578,125 ways in which the outcomes could occur. This problem is almost a little too big for a desk calculator and a little too small for a digital computer.

Summary

The extact multinomial test described in this article can be used to test the significance of variations in the numbers of observations distributed into two or more mutually-exclusive categories. When there are only two categories the test reduces to the binomial test. The test is valid for samples of any size but it quickly becomes prohibitively difficult to apply as the total number of observations or the number of categories increases. A comparison with the chi square test shows how seriously the latter may be in error when the number of observations is small.

REFERENCES

Cochran, W. G. The χ^2-test of goodness of fit. *Ann. math. Statist.*, 1952, **23**, 315–345.

Deininger, R. L. Human factors engineering studies of the design and use of push-button telephone sets. *Bell Sys. tech. J.*, 1960, **39**, 995–1012.

Kendall, M. G. *The advanced theory of statistics.* Vol. 1. London: Charles Griffin, 1947.

Pearson, E. S., & Hartley, H. O. (Eds.) *Biometrika tables for statisticians.* Vol. 1. Cambridge: Univer. Press, 1956.

Smith, J. G., & Duncan, A. J. *Sampling statistics and applications.* New York: McGraw-Hill, 1945.

Tate, M. W., & Clelland, R. C. *Nonparametric and shortcut statistics.* Danville, Ill.: Interstate Printers and Publishers, 1957.

(Received March 1, 1961)

Psychological Bulletin
1962, Vol. 59, No. 4, 311–319

THE ANALYSIS OF PROFILE DATA

JUM NUNNALLY
Vanderbilt University

During the last 20 years, the crystal-ball-gazing test interpreter has gradually been supplanted by the profile-gazing tester. He gazes steadfastly at the ups and downs on the profile chart for the Kuder Preference Record, the MMPI, and the Wechsler subtests; and from these he gives vocational advice, classifies the mentally ill, and searches for brain damage. Also, profile analysis has invaded psychological research, where comparisons are made between self and ideal-self ratings, and measures are made of interpersonal perception. Being scientific folk, some psychologists reasoned that if profile analysis is used (later it will be argued that sometimes it is better not to), then it should be used "objectively," i.e., in a mathematical and statistical framework.

There are three kinds of questions that profile analysis needs to answer:

1. How do you measure the relative similarity of two profiles?

2. How do you discriminate the typical profiles of two or more groups, e.g., MMPI profiles of different diagnostic groups?

3. How do you "cluster" profiles into homogeneous groups? A surprising amount of controversy has raged over how to answer these questions (see Haggard, Chapman, Isaacs, & Dickman, 1959, for some proposed solutions; see Sawrey, Keller, & Conger, 1960, for other proposed solutions, and a review of the relevant literature)—surprising because there are relatively simple, and satisfactory (we hope) answers to all three. The proposed answers to each of these will be discussed in turn.

Similarity of Profiles

There are two principal criteria by which to judge any measure of relationship: it should consider all of the information relevant to the comparisons and it should have mathematical properties which permit powerful methods of analysis. The first is partly a matter of preference; but once the desired measure is formulated, it greatly influences the kinds of analyses that can be performed.

Cronbach and Gleser (1953) reviewed the many proposed measures of profile similarity, criticized most of them, and recommended the use of the *d* measure, which is the square root of the sum of squared differences between profile elements. In an earlier paper, Osgood and Suci (1952) had also proposed the use of *d*. The argument for the use of *d* is that it considers all of the possible information in the profiles: *level*, *shape*, and *dispersion*. With respect to the first criterion given above for choosing a measure, *d* is appealing; and no one has proposed a more appealing measure.

The *d* measure also stands up well with respect to the second criterion. By using a measure of interpoint distance in Euclidean space, powerful methods of analysis are indeed available, ones which will be discussed more fully in the following sections. Because of these reasons, the author recommends, as others have, that profiles be considered as points in Euclidean space; however, as will be shown later, it is actually better to use a function of *d* rather than *d* itself in the analysis of profile data.

The use of d is appealing if, and only if, it is intended to compare profiles simultaneously with respect to level, shape, and dispersion. Later in the article it will be argued that in some studies it would be more meaningful to equate all profiles for level, and in other studies to equate for both level and dispersion, in which cases it would be more appropriate to use covariances, and correlations, respectively, rather than d. It will be shown that the same, powerful methods can be used with "raw" profiles as can be used with covariances and correlations.

Discrimination of Groups

If one accepts the Euclidean model, a powerful method of analysis is available for discriminating the typical profiles of two or more groups, namely, the linear multiple-discriminant function (Tatsuoka & Tiedeman, 1954). This will provide the best (in a least-squares sense) linear combination(s) for discriminating the groups, and it offers a procedure for assigning new individuals to one of the groups. For example, the discriminant function could be applied to the problem of differentiating the MMPI profiles of paranoids, psychopaths, and schizophrenics; and the results could be used to classify new cases into one of the three groups.

Clustering "Raw" Profiles

Clustering raw profiles is the problem that has aroused so much discussion, and the major purpose of this article is to attempt a satisfactory solution. This solution probably would have been adopted long ago had it not been for one mistaken notion among some psychologists about multivariate analysis.

Let us set the problem in focus by imagining that we are studying

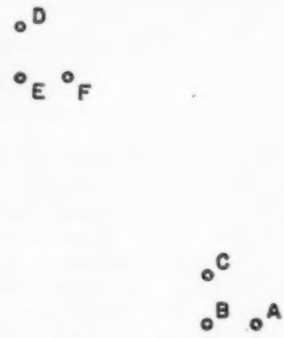

FIG. 1. Interpoint distances for six persons.

MMPI profiles and that we have the profiles from a broad sample of psychotic patients. We want to study the interrelations among the raw profiles in such a way as to say how many "kinds" (clusters) of profiles there are, and we want to measure the extent to which each patient belongs to each cluster. First, we will assume that relationships among the profiles should be pictured as interpoint distances in Euclidean space. (Some arguments for so doing were given above.)

In Figure 1 are pictured the hypothetical points for six patients, which are shown as lying in a two-space in order to simplify the illustration. By arbitrarily designating the distance from Person a to Person b as 1, all of the interpoint distances are set, and these are presented in Table 1.

In looking at Figure 1 and Table 1 it is obvious that there are two clustters, defined, respectively, by patients a, b, and c, and by patients d, e, and f. If in actual research there were so few cases involved and such definite clusters were present, no refined method of analysis would be needed; but this is almost never the case. A method of analysis will be

demonstrated which can recover these clusters and can be used equally well with any number of cases and regardless of the relative "visibility" of clusters.

It is apparently not widely known that *d* matrices such as that in Table 1 can be factored. The method was derived by G. J. Suci (Osgood, Suci, & Tannenbaum, 1957). Suci and I cooperatively explored his method of factoring *d* and found it to be a special case of raw score factor analysis. This is where the major misconception arises: some psychologists are evidently unaware that raw score cross-products can be factored in the same way as correlation coefficients are factored.

The failure to realize that factor analysis is not restricted to correlation coefficients is either directly evident or implied in many of the papers relating to methods of clustering profiles. Here is an example (Sawrey et al., 1960):

> Surely all factor analytic studies have not been interested in shape alone, yet this is, in fact, all that correlations, and *consequently* [italics added] factor analysis, takes into account (p. 670).

An Example of Raw Score Factor Analysis

Because of the unfamiliarity of factoring raw score cross-products, a worked-out example will be given. The first step is to obtain the sum of raw cross-products for each pair of patients over the profile elements. For the MMPI this consists of accumulatively multiplying the scores on corresponding scales for each pair of patients. A hypothetical matrix of such cross-products corresponding to the *d* matrix in Table 1 is shown in Table 2. Because I have chosen an artifical example, the cross-product terms look different from what would be obtained from an actual study of MMPI profiles.

How should one analyze Table 2 in order to obtain clusters? The answer is to factor analyze, and any of the methods commonly used with correlations can be applied: square root, multiple group, centroid, principal components, or what not. In doing this, the customary formulas are applied in the customary ways. Let us see what a centroid analysis provides.

For the first factor, sum the elements in each column, find the square root of the sum of the column sums, and divide this into each of the column sums. These are loadings on the first centroid factor in the raw score space. Use the first factor loadings to obtain a first set of residuals, reflect, extract a second set of centroid loadings, and continue in this manner until residuals are "small" or until enough factors have been obtained to satisfy the experimenter's curiosity.

By choosing a set of points in a two-space, only two factors are needed to explain the cross-products, and, consequently, the second residuals differ from zero only by rounding errors. Also, as would necessarily be the case, the sums of squares of "loadings" in rows of the factor matrix are identical to the original diagonal ele-

TABLE 1

MATRIX OF *d*'S FOR POINTS SHOWN IN FIGURE 1

Person	Person					
	a	b	c	d	e	f
a	.0	1.0	1.4	7.8	7.1	6.4
b	1.0	.0	1.0	7.2	6.4	5.8
c	1.4	1.0	.0	6.4	5.7	5.0
d	7.8	7.2	6.4	.0	1.0	1.4
e	7.1	6.4	5.7	1.0	.0	1.0
f	6.4	5.8	5.0	1.4	1.0	.0

ments in the cross-product matrix (which are the sums of squared scores over the profile elements for each patient).

By applying the orthogonal transformation shown in Table 2, a rotated factor solution is obtained. The clusters shown in Figure 1 and Table 1 are clearly evidenced in the rotated factor solution, and the factor loadings tell how much each patient belongs to each factor. In Figure 2 the rotated factors are plotted, and the obtained set of interpoint distances is identical to that shown in Figure 1. If one wants to cluster profiles, raw score factor analysis is a powerful and directly applicable procedure.

How Raw Score Analysis Works

Elements in profiles (e.g., the Paranoid scale of the MMPI) can be considered as mutually orthogonal axes in Euclidaen space. Each profile can be "plotted" as a point in the space, and *d* measures the distance of points from one another.

TABLE 2

RAW SCORE CROSS-PRODUCTS AND FACTOR SOLUTION FOR POINTS SHOWN IN FIGURE 1

Person	Cross-products					
	Person					
	a	b	c	d	e	f
a	36	30	30	6	6	12
b	30	25	25	5	5	10
c	30	25	26	11	10	15
d	6	5	11	37	31	32
e	6	5	10	31	26	27
f	12	10	15	32	27	29
Column sums	120	100	117	122	105	125
First factor	4.58	3.81	4.46	4.65	4.00	4.77
	First residuals					
Person	a	b	c	d	e	f
a	15.02	12.55	9.57	−15.30	−12.32	− 9.85
b	12.55	10.48	8.01	−12.72	−10.24	− 8.17
c	9.57	8.01	6.11	− 9.74	− 7.84	− 6.27
d	−15.30	−12.72	− 9.74	15.38	12.40	9.82
e	−12.32	−10.24	− 7.84	12.40	10.00	7.92
f	− 9.85	− 8.17	− 6.27	9.82	7.92	6.25
Column sums after reflexion	74.61	62.17	47.54	75.36	60.72	48.28
Second factor	− 3.89	− 3.24	− 2.48	3.92	3.16	2.51
	Second residuals					
Person	a	b	c	d	e	f
a	−.11	−.05	−.08	−.05	−.03	−.09
b	−.05	−.02	−.03	−.02	.00	−.05
c	−.08	−.03	−.04	−.02	.00	−.05
d	−.05	−.02	−.02	−.07	.01	−.02
e	−.03	.00	.00	.01	.01	−.01
f	−.09	−.05	−.05	−.02	−.01	−.05

TABLE 2—Continued

Person	Centroid factors		
	I	II	h^2
a	4.58	−3.89	36
b	3.81	−3.24	25
c	4.46	−2.48	26
d	4.65	3.92	37
e	4.00	3.16	26
f	4.77	2.51	29
	Transformation matrix		
	A	B	
I	.763	.647	
II	−.647	.763	
	Rotated factors		
Person	A	B	
a	6.01	.00	
b	5.00	−.01	
c	5.01	1.00	
d	1.01	6.00	
e	1.01	5.00	
f	2.02	5.00	

Raw score factor analysis provides a basis (or semibasis) for the profile space. Because any sufficient basis preserves distances between points, the factor loadings preserve the original *d*'s. In the example worked out previously, this can be tested by obtaining *d*'s from the two rotated factors. This shows, for example, that the *d* between Persons a and b is, within the limits of rounding errors, 1, which is what was given in Table 1. Similarly, all of the *d*'s can be calculated from the factor matrix. If factoring is not complete, then the factor matrix will serve to explain the bulk of the original distances.

The difficulty with most of the proposed measures of profile similarity is that they are non-Gramian,[1] e.g., Cattell's r_p (1949), and, consequently powerful methods of multivariate analysis cannot be used. Of course, matrices of cross-products are necessarily Gramian, and powerful methods of multivariate analysis, such as

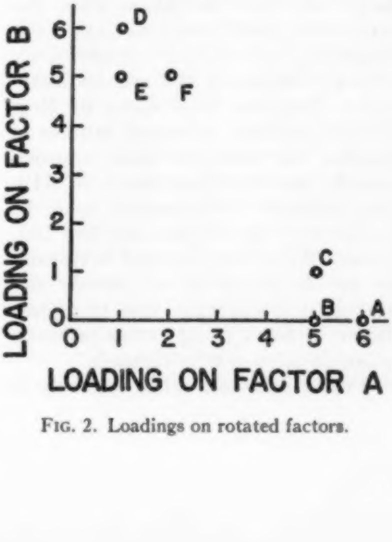

FIG. 2. Loadings on rotated factors.

[1] By definition a Gramian matrix is one whose elements consist of cross-products (Hohn, 1958, p. 202).

factor analysis, can be applied to them. Whenever there is a choice between a number of descriptive measures where one is Gramian and the others are not (e.g., the choice between point biserial and biserial correlation), it greatly facilitates the analysis of results to choose the Gramian measure.

Preparation for Analysis

Much of the controversy about the analysis of profile data has concerned what, if any, transformations should be made before the data is analyzed. Regardless of what transformations are made, factor analysis of cross-product terms is a powerful method available to search for clusters. Two kinds of transformations have been proposed: transformations of distributions of individual differences on profile elements, and transformations of profiles as a function of intra-individual distributions. Each will be considered in turn.

Individual differences. If the individual profile elements have grossly different standard deviations, some elements will contribute more to the interpoint distances than will others. For example, on the Rorschach, the number of $F+$ responses has a much larger standard deviation than the number of pure C responses, and, consequently, the former would more strongly influence the size of interpoint distances in a space of Rorschach profiles. In many studies of profiles the elements have approximately the same dispersions: MMPI, the Semantic Differential, and the subtests of the multifactor test batteries. When the standard deviations of profile elements are grossly different, it is generally wise to equate them before using cross-products analysis to search for clusters.

Profile elements differ not only in terms of dispersions, they also differ in terms of their factor compositions. For example, profile elements on the Semantic Differential differ in terms of the factors of *evaluation*, *potency*, *activity*, and others. To the extent that one factor is more prominent than others in the collection of profile elements, that factor will more strongly influence the size of interpoint distances. One way to offset the differential influences of such factors is to factor analyze the profile elements (correlating over persons) and reduce the profiles to sets of factor scores. Then the cross-products analysis of profiles can be made of the sets of factor scores. For example, instead of beginning with a space of interpoint distances formed by individual Semantic Differential scales, we can begin the analysis of profiles by constructing a space of Semantic Differential factor scores. Factor analysis of cross-products applies equally well in this situation, and it can be used regardless of the kinds of transformations that are made on profile elements.

If the purpose of the analysis is to discriminate the typical profiles of two or more groups (Question 2, posed earlier), then nothing can be gained from transforming score distributions on profile elements. The discriminant function will provide the same results whether or not the dispersions are equated. Also, the resolution of profile elements into factors cannot possibly add to the discriminability that would be obtained from a discriminant-function analysis of the elements themselves. However, a prior factor analysis of profile elements is sometimes wise because it simplifies the subsequent use of the discriminant function, leaves less opportunity for the discriminant function to "take advan-

tage of chance," and usually makes the discriminant functions more interpretable.

Intraindividual distributions. If, as some claim, profiles should be clustered with simultaneous respect to level, shape, and dispersion, then factor analysis should be made of raw cross-products, either on the untransformed profile elements or after transformations of the kinds discussed previously are made.

If level is considered to be unimportant in clustering profiles, then the means of all profiles should be equated before the analysis, preferably equated to zero. Next form cross-product terms and divide each by the number of profile elements; then factor by any of the conventional methods. This is called covariance factor analysis, but it is only a special case of cross-products analysis.

If both level and dispersion are considered unimportant, convert all profiles to standard scores, standardizing over the profile elements. Then form a matrix of cross-products and divide each term by the number of profile elements. This gives a correlation matrix, and no one needs to be told that it can be factor analyzed.

If the purpose of the analysis is to discriminate the typical profiles of two or more groups (Question 2 posed earlier), it is an empirical question whether or not transformations of intraindividual distributions will help or hinder the outcome in particular studies. For example, if in a particular study all of the profiles are equated for *level*, this might increase discriminability or it equally well may lower discriminability. Consequently, before applying the discriminant function, it is wise to compare groups with respect to level, shape, and dispersion. If groups differ inconsequentially on any of the components, it is wise to remove that component(s) before applying the discriminant function.

Should Profiles Be Analyzed?

Most of this article is concerned with *how* to analyze profile data. Equally important is the initial decision in research to make comparisons among score profiles. Perhaps in many situations it would be wiser not to make such comparisons at all.

The decision to use profile analysis is determined in part by preferences for methodologies, which are, in essense, wagers about the likely research payoff in the long run from choosing one method of investigation rather than another. The reader can judge for himself whether the studies using comparisons of profiles (e.g., measures of "assumed similarity" in interpersonal perception) have borne the expected fruit.

If analyses are made of the relations among raw profiles, in which level, shape, and dispersion are preserved, the results are often difficult to interpret. Particular results may be due to any one of the three profile components, and, without reanalyzing differently, there is no way to unravel the puzzle. Even those who initially advocated the analysis of raw profiles have since either advocated or practiced separate analyses of level, shape, and dispersion (for example, Cronbach, 1958).

It was argued that factor analysis of cross-products is the best way to cluster profiles. However, when such analyses are made of raw profiles, it is sometimes difficult to interpret the results. Most of us have become so familiar with correlation coefficients, and factor analyses of them, that it raises some anxiety to look at factor loadings like -68.21 and 4.89.

A good argument can be given for

the use of profile analysis in studies of personnel decisions, e.g., selecting men for a particular job, or classifying patients for different kinds of treatment. If criterion variables are available, the validity of any decision strategy based on profile analysis can be determined directly. Then the only sense in which it is necessary to justify the analysis of profiles is to show that it works better than some other approach. For example, it might be found that a discriminant-function analysis of score profiles is more effective in classifying mental patients than is a multiple regression approach. The major difficulty in validating profile analyses is that in many types of personnel decisions there are no adequate criteria available, and the questions of whether to use profile analysis and, if so, how, are left moot.

It is more difficult to argue for the use of profile comparisons in testing psychological theories. Many efforts have been made to assert hypotheses about interpersonal perception, psychotherapy, empathy, and others, in terms of profile similarities and differences. A major drawback to formulating such hypotheses is that they inevitably involve the semi-undefinable quality of "similarity." Also, the general experience has been that the results of such studies often are much clearer when univariate comparisons rather than profile comparisons are made. This is illustrated in some of the studies of before-after therapy comparisons of profiles of "self" and "ideal-self" ratings. What has generally been found is that nearly everyone has the same "ideal" before therapy, and the ideal changes little during therapy. The change, if any, is in the self, and the change is mainly toward higher self-esteem (Rogers & Dymond, 1954, p. 417). Consequently, rather than assert vague and complex hypotheses about similarities among profiles before and after, it is much more meaningful to hypothesize that successful therapy raises self-esteem. Studies of interpersonal perceptions (for example, Bass & Fiedler, 1959) have also indicated that univariate comparisons often are more revealing than profile comparisons.

Summary

Methods were suggested for handling three problems in the analysis of test profiles: measuring the similarity of profiles, discriminating the typical profiles of two or more groups, and clustering profiles into homogeneous groups. The suggested methods were, respectively: picturing profiles as interpoint distances in Euclidean space, use of the linear multiple-discriminant function, and factor analysis of profile cross-product terms. Some suggestions were given about transformations of profile data before further analysis. Some opinions were stated about the appropriateness of profile analysis in studies of personnel decisions and in investigations of psychological theories.

REFERENCES

Bass, A. R., & Fiedler, F. E. Interpersonal perception scores: A comparison of *D* scores and their components. (Tech. Rep. No. 5) Urbana: Univer. Illinois, Group Effectiveness Research Laboratory, 1959. (Mimeo)

Cattell, R. B. r_p and other coefficients of pattern similarity. *Pyschometrika*, 1949, **14**, 279–298.

Cronbach, L. J. Proposals leading to analytic treatment of social perception scores. In R. Tagiuri and L. Petrullo (Eds.), *Person*

perception and interpersonal behavior. Stanford: Stanford Univer. Press, 1958. Pp. 353–379.

Cronbach, L. J., & Gleser, G. C. Assessing similarity between profiles. *Psychol. Bull.*, 1953, **50**, 456–473.

Haggard, E. A., Chapman, Jean P., Isaacs, K. S., & Dickman, K. W. Intraclass correlation versus factor analytic techniques for determining groups of profiles. *Psychol. Bull.*, 1959, **56**, 48–57.

Hohn, F. E. *Elmentary matrix algebra*. New York: Macmillan, 1958.

Osgood, C. E., & Suci, G. J. A measure of relation determined by both mean differences and profile information. *Psychol. Bull.*, 1952, **49**, 251–262.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning*. Urbana: Univer. Illinois Press, 1957.

Rogers, C. R., & Dymond, Rosalind F. (Eds.) *Psychotherapy and personality change*. Chicago: Univer. Chicago Press, 1954.

Sawrey, W. L., Keller, L., & Conger, J. J. An objective method of grouping profiles by distance functions and its relation to factor analysis. *Educ. psychol. Measmt.*, 1960, **20**, 651–673.

Tatsuoka, M. M., & Tiedeman, D. V. Discriminant analysis. *Rev. educ. Res.*, 1954, **24**, 402–420.

(Received March 1, 1961)

Psychological Bulletin
1962, Vol. 59, No. 4, 320-328

ON SIMPLE METHODS OF SCORING TRACKING ERROR[1]

E. C. POULTON
Applied Psychology Research Unit, Cambridge, England

The thesis of this paper is that simple measures of the error for one-dimensional tracking, provided the right ones are used, can reveal the response strategy which the subject (*S*) adopts without involving an inordinate amount of work scoring records. The measures are the mean constant error (*CE*), and the standard deviation (*SD*) of the error indicating within-*S* variability, computed separately for position and time at various points on the input. Samples of the measures can be obtained relatively easily and quickly by hand from a record like that in Figure 1. When an electronic display is used it is possible to produce such a record by feeding the input and response into two channels of an oscillographic recorder, and subsequently superimposing them.

No special merit is attached to measuring by hand. Once it has been decided which are the best measures to use, it is possible either to build electronic devices to do the measuring, or to record performance on for example magnetic tape, and to feed the tape into a computer which is first programed to produce one measure, and then programed to produce another (Webber & Adams, 1960). However, electronic devices and computer programs only answer the questions which the experimenter (*E*) asks of them; they cannot tell him what new questions to ask. Unless *E* has available almost unlimited resources for automatic data processing so that he can test out quite unlikely hypotheses, it may be advisable to make measurements by hand on sample records in order to be sure of not missing new and unexpected features. A simple clinical assessment from watching *S* perform or from serving as *S* may motivate *E* to make the necessary analyses, but is unlikely to tell him precisely what are the best measurements to make.

The parallel approach using the describing functions of the engineer can be dismissed in a few words, since it is less relevant to psychology, and has been ably summarized by McRuer and Krendel (1957). Describing functions are based upon mathematical systems of analysis designed primarily to determine the numerical values of the parameters of servomechanisms. They are thus capable of giving exact numerical values to those aspects of human tracking performance which resemble the parameters of servomechanisms (Ellson, 1959). But they are not so suited as simple measures are for determining the details of the ways in which human operators do not behave like servomechanisms; these include most of the phenomena studied by psychologists (Adams, 1961, pp. 56–60).

Some Simple Methods of Scoring

Overall Error in Position

One of the simplest measures of performance is the mean error in position neglecting the sign. During World War II Craik and Vince

[1] The author is indebted to M. Stone for discussions on the statistical aspects of these problems, to J. A. Adams and E. J. Archer for constructive criticisms, and to the British Medical Research Council for financial support.

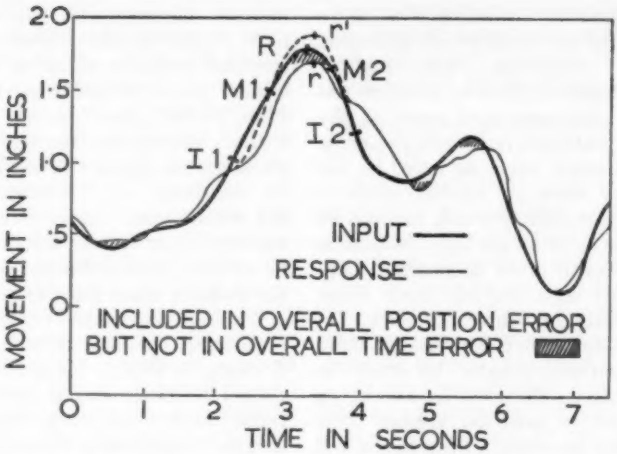

FIG. 1. Section of a record from pursuit tracking in one dimension with a preview of 2.5 seconds. (R, r, and r[1]: points at which the input, the undershooting response, and a comparable overshooting response represented by the broken line, reversed direction, I_1 and I_2: points of inflection on the input. M_1 and M_2 lie on the input half way in time between R and I_1 or I_2.)

(1943, p. 5) recorded performance in one-dimensional tracking on a smoked drum moving at 50 millimeters per minute, measured the error in position every 1.0 millimeter (1.2 seconds), and took the mean value. This is equivalent to drawing in Figure 1 a series of vertical lines connecting the two functions and computing their mean height. (If a search is to be made for high frequency components in the response, it is necessary to sample at a frequency at least double that of the highest component. But when using simple methods of scoring there is little point in sampling the error more frequently than once per second provided an adequate length of record is available. For the only valid statistical test of whether the results are representative of the population from which the *S*s are drawn depends upon the differences between *S*s; and there is a stage beyond which increasing the reliability of the means of individual *S*s makes little difference to the variability between *S*s).

If it is desired to increase the penalty upon large errors, a function of squared error such as mean squared error or root mean squared error may be used instead of mean error. If small errors do not matter at all, a "target" area of a selected size may be used and error be scored only beyond it. This is equivalent to increasing the width of the input function in Figure 1, and measuring the error from its edges. The method is somewhat analogous to measuring only the time for which *S* is outside the designated target area (see time on target below).

A major disadvantage of using only a measure of the overall error in position neglecting the sign, is that it confounds what are probably true errors of position with errors of timing. The heights of the shaded areas in Figure 1 can be looked upon as representing true errors in position,

since S should have stopped when he reached the points at which the input reversed direction, but instead stopped short of them or went on too far. In contrast, over most of the distance between reversals the error can be looked upon as more in the nature of error in timing, since S covered the right ground, but did so either too early or too late. Measures of the overall error in position neglecting its sign include both these rather different sources of error.

A measure which is probably more or less uncontaminated by errors in timing can be obtained by averaging the errors in position taking their signs into account. This mean *CE* shows the extent to which the response is on average to one side or other of the input. Unfortunately it gives no indication of whether S tended to overshoot or undershoot at reversals, since adjacent overshoots tend to cancel out in the side-to-side dimension, and the same applies to adjacent undershoots. The mean *CE* is not very contaminated by errors in timing provided the input is on average symmetrical with respect to time and position (as harmonic inputs are), and provided the same is approximately true of the response. For under these conditions the *CE*s for time tend to cancel out in the position dimension, and vice versa.

Overall Error in Time

During World War II Helson (1949, p. 477) used measures of time error in one-dimensional tracking as well as measures of position error. Measuring the error in time gives a mean lag if the sign of the error is taken into account, in addition to a mean error when the sign is neglected. In pursuit tracking with a reasonably random harmonic input of high frequency which cannot be seen in advance, the two measures tend to give similar values since the response is rarely ahead of the input under these conditions (Poulton, 1954, Table 3, fast course). However with a random low frequency or simple harmonic input the mean lag may be relatively small compared with the mean error. Again a function of squared error can be used instead of mean error if it is desired to increase the penalty upon large errors.

Taking the mean error in time neglecting the sign is equivalent to drawing in Figure 1 a series of horizontal lines connecting the two functions, and computing their mean length. Integrating the error in time with respect to position in this way gives the same result as integrating the error in position with respect to time, except in so far as S overshoots or undershoots at the reversals in the direction of movement of the input. Figure 1 shows that overshoots are not taken into account in the time dimension, since there is no input to which they correspond. Similarly undershoots leave a loop of the input without a corresponding part of the response function. Thus the shaded areas in Figure 1, which can be looked upon as predominantly error in position, are included in the overall error in position, but not in the overall error in time.

Just before an undershot reversal such as R in Figure 1, S is typically much behind the input, whereas just after the reversal he is typically much ahead of it. Conversely in overshooting, which is represented by the broken line in Figure 1, S is typically first ahead of the input and then behind. The sign of the change in time error introduced by an undershoot or overshoot before a reversal is the opposite of the sign of the change introduced after the re-

versal. Thus if the sign of the time error is taken into account, as in calculating the overall lag, the change in the error just before the reversal tends to cancel the change just after the reversal. Mean lag is thus not appreciably affected by overshooting or undershooting. This is not the case for the mean error in time neglecting the sign, unless *S* consistently lags further behind the input than the sizes of the changes in time error introduced by overshoots and undershoots. It can never be the case for overall measures involving squared time errors, since $(L+c)^2+(L-c)^2 \geq 2L^2$; these measures are necessarily inflated by overshooting and undershooting, even though they exclude the shaded areas in Figure 1.

The failure to take account of the shaded areas in Figure 1 is a major disadvantage of using the overall error in time with sign neglected as the sole measure of performance in tracking inputs which reverse direction. Time on target does not meet this particular difficulty. This measure corresponds to increasing the thickness of the input function in Figure 1 to the width of the target area, and measuring the time for which the response line lies within its boundaries. However time on target takes no account of the size of the excursions from the target area, and its exact meaning has been questioned by Bahrick, Fitts, and Briggs (1957) on these and other grounds.

Cross-correlation is a more sophisticated technique for determining the average time relationships between the input and response. This involves correlating the two after moving the response by each of a number of fixed steps along the time dimension (Merrill & Bennett, 1956). The size of step which the response has to be moved forward or backward along the time dimension in order to give the largest correlation with the input indicates *S*'s overall lag or lead with respect to the input.

Average Error at Particular Points on the Input

With harmonic inputs average errors can be calculated at particular points such as reversals in direction, points of inflection, and points half way in time between reversals and points of inflection. Figure 1 shows that between reversals it is often impossible to specify with any degree of certainty the corresponding points on the wiggly response record; thus the mean lag computed as described above is probably the most meaningful measure here. At the points of inflection (I_1 and I_2 in Figure 1) the mean lag is probably more or less uncontaminated by errors in positioning, since these points are placed symmetrically on the input. However at the points half way in time between reversals and points of inflection (M_1 and M_2 in Figure 1) a tendency to undershoot increases the mean lag before reversals and correspondingly reduces it afterwards. A tendency to overshoot typically has an opposite though smaller effect. To the extent that the effect of overshooting does not fully balance that of undershooting, simple variability in the overshoot-undershoot dimension should have an effect similar to undershooting, though less marked.

A reversal on the input (Point R in Figure 1) can be compared directly with the corresponding reversal on the response function (Point r or r^1). At these points it is thus possible to obtain four separate measures: a *CE* and an *SD* of error related to the position of the response irrespective of its timing, and two similar measures related to the timing of the re-

sponse without regard to its position. The sizes of the overshoots and undershoots (the heights of the shaded areas in Figure 1) probably provide the most useful measures in the position dimension. An alternative version of the *CE* in position, which shows only the extent to which the response is on average too far to one side or other of the input, is less relevant to the primary problem facing *S*.

An Illustration

Table 1 shows results from an as yet unpublished experiment to illustrate some of the more useful simple methods of scoring. Each entry in the table is based upon only 120 measurements, 10 from each of 12 *S*s. The complete data for the Preview version thus required only 840 measurements, and the same is true of the Slit version. This has been done deliberately in order to minimize labor, and to show how much can be discovered in spite of this. More generous sampling calls if possible for automatic methods of scoring.

METHOD

Apparatus. An irregular curved input, of which Figure 1 shows a sample, was drawn on a paper tape which moved towards *S* at a rate of 1.0 inch per second. The frequencies in the input were 26 cycles per minute, 21 cycles per minute, and a component of 10.5 cycles per minute which had twice the amplitude of the other two. The maximum amplitude of movement of the input was 1.75 inches. A ball-point pen was used as a stylus. The stylus could be moved in a slit lying over the paper tape at right angles to it.

Task. For the data in Table 1 *S* had to keep the stylus on the input by moving it from side to side in the slit. In the Preview version he could see the input 2.5 seconds (2.5 inches) ahead of the stylus, as a walker can normally see the footpath ahead. In the Slit version he could see the input only in the slit in which the stylus moved. The slit had a width of .1 inch.

Procedure. Each trial lasted 30 seconds. Half the *S*s did the Preview version first, and half the Slit version. Data was also collected when a gap separated the input from the stylus, and *S* had to keep the two aligned, but it is not shown in Table 1. Practice was deliberately restricted, so that for the results

TABLE 1

SOME SIMPLE METHODS OF SCORING ERROR

Points at which error measured	Preview version				Slit version			
	CE		*SD*		*CE*		*SD*	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Error in position (mm.)								
Overall sample	.0	.13	1.61[ab]	.16	.12R	.26	3.57[ab]	.31
Reversals	.03L	.15			.38L	.32		
			1.05[ab]	.11			2.41[ab]	.25
	.41U[b]	.10			1.13O[b]	.36		
Error in time (sec.)								
Overall sample	.039[b]	.013	.117[b]	.010	.11[b]	.015	.15[ab]	.013
Reversals	.064[bc]	.011	.101[bc]	.008	.11[b]	.016	.12[ab]	.008
¼ cycle after reversals	.011[b]	.009	.082[b]	.006	.14[b]	.015	.13[b]	.014
Points of inflection	.031[bc]	.013	.074[bc]	.006	.13[b]	.016	.12[b]	.014
¼ cycle before reversals	.082[b]	.013	.096	.010	.12[b]	.017	.12	.015

Note.—L and R indicate that the response was too far to the left or right, while U and O indicate a tendency to undershoot or overshoot. The mean *CE*s for time were all lags.

[a] Overall sample—Reversals $p < .05$ or better.
[b] Preview—Slit $p < .05$ or better.
[c] Reversals—Points of inflection $p < .05$ or better.

in the table each S had performed for altogether only between 2.0 and 5.0 minutes on each version. The amounts of practice on each version were counterbalanced between Ss.

Subjects. These were 12 young enlisted men in the British Royal Navy, none of whom had done much tracking.

Scoring. Each mean in the table is based upon 10 measurements from the record of each S. The measures at reversals, at points of inflection, and at the two series of points midway in time between reversals and points of inflection came from the second half of the 30-second trial. At the three latter sets of points on the input the time error corresponds to the horizontal distance in Figure 1 between the two functions. At reversals the time and position error are, respectively, the difference between when and where the input reversed direction and when and where the response did so. Where S stopped for an appreciable time before moving off again in the opposite direction, as near r in Figure 1, the time error is computed from the average of the time at which he stopped and the time at which he moved off again.

The measures in the overall sample represent performance averaged over all points on the input. The sample is based upon 10 points separated by 1.6 seconds, also from the second half of the trial. (This periodicity is not related in a simple way to any of the three frequency components of the input, and makes the sample cover approximately the same length of record as the samples at particular points on the input.) Where an error in time could not be measured at the selected point because S had undershot, the next point was taken instead.

Calculations. The means *SD*s show the variability within Ss, the *SE*s the variability between Ss. Thus the *SE* of an *SD* indicates the size of the individual differences in variability. The reliability of the differences between means was assessed by two-sample t tests with 11 degrees of freedom, using two tails.

Results and Discussion

It has been suggested above that performance at reversals probably provides the most useful measures of the errors in position. Table 1 shows that in the Preview version S tended to undershoot by an average of .41 millimeters ($p < .002$), whereas in the Slit version he tended to overshoot by an average of 1.13 millimeters ($p < .01$). The *SD*s of the errors in position at reversals were over twice as large in the Slit version as with Preview ($p < .001$). In the overall sample, which shows performance averaged over all points on the input, the mean *CE* in position is unlikely to be very contaminated by time error, but the *SD* of the error in position necessarily contains an unknown component of time error. There was no significant difference between the overall sample and reversals in the extent to which the response was on average to one side or other of the input ($p < .05$), but as expected the *SD*s of the overall sample were significantly too large in both versions as compared with the *SD*s at reversals ($p < .01$).

As already indicated, the mean *CE*s in time at reversals are not contaminated by errors in positioning, and the mean *CE*s of the overall sample and at the points of inflection are unlikely to be very contaminated. The combined means of the *CE*s $\frac{1}{8}$ cycle before and after reversals are also unlikely to be very contaminated, but a tendency to undershoot, and even simple variability in positioning at reversals, is likely to increase the mean time lags before reversals, and to reduce correspondingly the mean lags after reversals. Table 1 shows that in the Preview version the mean time lag was twice as great at reversals as it was at the points of inflection on the input ($p < .05$). The average time lag at the intermediate points on the input $\frac{1}{8}$ cycle before and after revesals, .47 seconds, lay intermediately. Thus in the Preview version S did not simply reproduce the input as accurately as he could with a constant time lag; his timing varied significantly at different points on the input cycle.

The nature of this nonlinear relationship would not have been so easy to determine using the describing functions of the engineer, since servomechanisms do not normally behave like this (see the introduction).

In the Slit version there were no significant differences between the mean time lags at different points on the input ($p > .05$). If differences exist, larger samples of data are required in order to reveal them. None of the *SD*s of the error in time is an adequate measure of the overall variability in the time dimension. The *SD*s at reversals are likely to be smaller than the true overall variability in timing, since only one point on the input is represented. However, unlike the remaining *SD*s in time in Table 1, the *SD*s at reversals are not contaminated by variability in the position dimension; thus it is possible to make a valid comparison of the variability in timing at these points between the two versions. Table 1 shows that there was significantly more variability in timing at reversals in the Slit version ($p < .01$), although the size of the difference was not large.

Two Different Strategies

The effect of size of preview was investigated in a previous experiment using the same input and apparatus as used here. Overall performance was found to change markedly when the preview was increased from zero (as in the Slit version) to .4 second; but there was no significant further change when the preview was extended from .4 to 8.0 seconds (Poulton, 1954, p. 406). The 2.5-second preview used here was chosen to be well on that part of the function where overall performance had ceased to change appreciably with increase in preview. The differences between the Slit and Preview versions in Table 1 can thus be taken to represent the maximum effect which a change in preview is likely to produce. By comparing the two versions it is possible to indicate the response strategy adopted in each case, and thus to demonstrate the usefulness of the simple measures given in the table.

In the Preview version Table 1 shows that the mean *SD* of the error in position at reversals was less than half the size of that in the Slit version. In addition the mean lag was twice as large at reversals as at the points of inflection on the input, whereas there was little difference in the Slit version. Thus in attempting to keep as much as possible on the input, *S* adopted the strategy in the Preview version of minimizing overshoots and undershoots at reversals by approaching them more slowly than did the input itself, and catching up again at the start of the return movement.

The stimulus conditions which presumably determined this strategy were as follows: close to the points of inflection on the input a small error in timing produced a considerable misalignment, since here the input was moving at its maximum velocity (see Figure 1). In contrast, close to reversals even a relatively large error in timing did not produce much misalignment provided the amplitude of the response was correct, since the input was more or less stationary; whereas an overshoot or undershoot not only produced a misalignment proportional to its size, but the misalignment tended to remain for quite a time, since both input and response moved so slowly here (see Figure 1). Misalignment was thus minimized by concentrating upon correct timing near the points of inflection, and upon correct positioning near reversals.

In the Slit version Table 1 shows

that the *SD* of the error in position at reversals was over twice as large as in the Preview version. Also *S* overshot by an average of 1.13 millimeters, as compared with a mean undershoot of only about one third the size in the Preview version. In the time dimension the mean lag in the overall sample was almost three times as large in the Slit version as with Preview, although it was by no means as long as a visual reaction time (RT) which is usually given as about .18 second (Woodworth, 1938, p. 324). In addition, at least at reversals, the timing was rather more variable in the Slit version, although there was less difference than in the Preview version between the mean lag at one point on the input and another.

The Slit version presented what was effectively an insoluble problem: *S* had to keep up with an irregular input which he could not see in advance. In so far as he attempted to compensate for his RT he had therefore to act on his predictions as to what the input was about to do, and thus to risk overshooting when the input stopped and reversed direction unexpectedly, and undershooting when the input went on further than he expected. In a typical RT experiment his behavior would produce so-called "premature" or "false" reactions. Faced with this problem, *S* adopted a strategy which was a compromise between on the one hand keeping up with the input regardless of overshooting and undershooting at reversals, and on the other hand of minimizing overshoots and undershoots by following a full RT behind the input.

Summary

For one-dimensional tracking simple measures are described which can be made in terms of both position and time. The measures may be averaged over all points on the input, or may be averaged for only one kind of point; e.g., at the reversals in direction, or at the points of inflection on the input. At reversals it is possible to score the error between the corresponding points on the input and response functions, and thus to produce measures of error in positioning which are uncontaminated by errors in timing and vice versa.

Overshoots and undershoots are probably the most relevant errors of positioning. The extent to which the response is on average to one side or other of the input, the overall lag or lead, and the mean lag or lead at the points of inflection on the input and at points situated symmetrically on either side of it, can all probably be computed in a reasonably uncontaminated form. Most other measures described confound more seriously errors of positioning with errors of timing.

Some unpublished data are used to illustrate the various measures. They show a previously unreported relationship which would not have been so easy to specify using the describing functions of the engineer. From the data it is possible to distinguish two different response strategies, which can be related to differences in stimulus conditions. The analysis demonstrates the increased insight into the stimuli influencing *S* in tracking, and into the strategies adopted, which can come from simple methods of scoring involving only a limited number of measurements.

REFERENCES

ADAMS, J. A. Human tracking behavior. *Psychol. Bull.*, 1961, **58**, 55–79.

BAHRICK, H. P., FITTS, P. M., & BRIGGS, G. E. Learning curves: Facts or artifacts? *Psychol. Bull.*, 1957, **54**, 256–268.

CRAIK, K. J. W., & VINCE, M. Psychological

and physiological aspects of control mechanisms with special reference to tank gunnery. Part 1. Report, 1943, Psychological Laboratory, Cambridge.

ELLSON, D. G. Linear frequency theory as behavior theory. In S. Koch (Ed.), *Psychology: A study of science.* Vol. 2. New York: McGraw-Hill, 1959.

HELSON, H. Design of equipment and optimal human operation. *Amer. J. Psychol.*, 1949, 62, 473–497.

MCRUER, D. T., & KRENDEL, E. S. Dynamic response of human operators. *USAF WADC tech. Rep.*, 1957, No. 56-524.

MERRILL, W. J., JR., & BENNETT, C. A. The application of temporal correlation techniques in psychology. *J. appl. Psychol.*, 1956, **40**, 272–280.

POULTON, E. C. Eye-hand span in simple serial tasks. *J. exp. Psychol.*, 1954, **47**, 403–410.

WEBBER, C. E., & ADAMS, J. A. Issues in the use of an analog-digital data system for the measurement of tracking behaviour. *USAF Off. Scient. Res. tech. Note*, 1960, No. 59-528.

WOODWORTH, R. S. *Experimental psychology.* New York: Holt, 1938.

(Received March 13, 1961)

Psychological Bulletin
1962, Vol. 59, No. 4, 329–343

THE PROCESS-REACTIVE CLASSIFICATION OF SCHIZOPHRENIA

WILLIAM G. HERRON
Saint Bonaventure University

The heterogeneity of schizophrenic patients and the lack of success in relating variable schizophrenic functioning to diagnostic subtypes (King, 1954) have indicated the serious limitations of the current neuropsychiatric classification of schizophrenia. In response to these limitations interest has arisen in a two-dimensional frame of reference for schizophrenia. Such a conception is based on the patient's life history and/or prognosis. A number of terms—malignant-benign, dementia praecox-schizophrenia, chronic-episodic, chronic-acute, typical-atypical, evolutionary-reactive, true-schizophreniform, process-reactive—have appeared in the literature describing these two syndromes. Process schizophrenia involves a long-term progressive deterioration of the adjustment pattern with little chance of recovery, while reactive schizophrenia indicates a good prognosis based on a history of generally adequate social development with notable stress precipitating the psychosis.

In view of the current favorable interest in this approach to the understanding of schizophrenia (Rabin & King, 1958) the present investigation is designed as an evaluative review of the literature on the process-reactive classification.

Early Prognostic Studies

The process-reactive distinction had its implicit origin in the work of Bleuler (1911). Prior to this the Kraepelinian influence had prevailed, with dementia praecox considered an incurable deteriorative disorder. Bleuler, while adhering to an organic etiology for schizophrenia, nonetheless observed that some cases recovered. This conclusion opened the field to a series of subsequent prognostic studies (Benjamin, 1946; Chase & Silverman, 1943; Hunt & Appel, 1936; Kant, 1940, 1941, 1944; Kretschmer, 1925; Langfeldt, 1951; Lewis, 1936, 1944; Malamud & Render, 1939; Mauz, 1930; Milici, 1939; Paskind & Brown, 1940; Wittman, 1941, 1944; Wittman & Steinberg, 1944a, 1944b) eventuating in formalized descriptions of the process and reactive syndromes in terms of specific criteria.

These early studies can be classified in three general catagories: studies correlating the outcome of a specific type of therapy with certain prognostic variables, studies descriptively evaluating prognostic criteria, and studies validating a prognostic scale.

The first category is illustrated by the attempt of Chase and Silverman (1943) to correlate the results of Metrazol and insulin shock therapy with prognosis, using 100 schizophrenic patients treated with Metrazol and 40 schizophrenic patients treated with insulin shock.

In the first part of this study the probable outcome of each of the 150 patients was estimated on the basis of prognostic criteria. The criteria considered of primary importance for a favorable prognosis were: short duration of illness, acute onset, obvious exogenic precipitating factors, early prominence of confusion, and atypical symptoms (marked by strong mixtures of manic-depressive,

psychogenic, and symptomatic trends), and minimal process symptoms (absence of depersonalization, derealization, massive primary persecutory ideas, and sensations of influence, conscious realization of personality disintegration, bizarre delusions and hallucinations, marked apathy, and dissociation of affect). When these conditions were reversed the prognosis was least favorable. The following factors were considered less important for a favorable prognosis: history of previous illness, pyknic body type, extrovert temperament and adequate prepsychotic life adjustment, catatonic and atypical subtypes. Asthenic body type, introversion, inadequacy of prepsychotic reactions to life situations, onset of illness after the age of 40, and hebephrenic and paranoid subtypes were considered indicative of unfavorable prognosis. Age of onset under 40, sex, education, and abilities, and hereditary background were not considered of prognostic importance. An analysis of the prognostically significant factors resulted in the evaluation of the prognosis for each case as good, fair, or poor.

Following termination of shock treatment all patients were followed-up for an average of 10 months and divided into three groups; much improved, improved, and unimproved. A comparison of the prognostic assessments with the results of shock indicated that of 43 cases in which the prognosis was considered good, 33 showed remissions, while of 74 cases with a poor prognosis, 63 did not improve. It was concluded that shock therapies were effective in cases of schizophrenia in which the prognosis was favorable, but were of little value when the prognosis was poor.

The second part of the research involved a reanalysis of the prognostic criteria in the light of the results of shock treatment. Short duration of illness and the absence of process symptoms were the most significant factors for favorable outcome, while long duration of illness (more than 2 years) and the presence of process symptoms were primary in determining poor prognosis.

A descriptive review of prognostic factors is seen in Kant's (1944) description of the benign (reactive) syndrome as cases in which clouding and confusion prevail, or in which the schizophrenic symptoms centered around manic-depressive features or cases with alternating states of excitement and stupor with fragmentation of mental activity. Malignant (process) cases are characterized by direct process symptoms. These include changes in the behavior leading to disorganization, dulling and autism, preceding the outbreak of overt psychosis. The most subtle manifestation of this is the typical schizophrenic thought disturbance. The patient experiences the process as a loss of normal feeling of personality activity and the start of experiencing a foreign influence applied to mind or body.

The third category includes the Elgin Prognostic Scale, constructed by Wittman (1941) to predict recovery in schizophrenia. It is comprised of 20 rating scales weighted according to prognostic importance: favorable factors are weighted negatively, and unfavorable factors are assigned positive weights. Initial validation involved 343 schizophrenic cases placed on shock treatment. Wittman and Steinberg (1944a) performed a follow-up study on 804 schizophrenics and 156 manic-depressive patients. The Elign scale proved effective in predicting the outcome of therapy in 80–85% of the cases in both studies, and has been

utilized in the work of Becker (1956, 1959), King (1958), and McDonough (1960) to distinguish the process-reactive syndrome. Included in the subscales of the Elgin scale are evaluations of prepsychotic personality, nature of onset, and typicality of the psychosis relative to Kraepelin's definition.

Studies with Detailed Process-Reactive Criteria

The synthesis of early studies is found in the research of Kantor, Wallner, and Winder (1953) establishing detailed criteria for distinguishing the two syndromes on the basis of case history material. A process patient would exhibit the following characteristics: early psychological trauma, severe or long physical illness, odd member of the family, school difficulties, family troubles paralleled by sudden changes in the patient's behavior, introverted behavior trends and interests, history of a breakdown of social, physical, and/or mental functioning, pathological siblings, overprotective or rejecting mother, rejecting father, lack of heterosexuality, insidious gradual onset of psychosis without pertinent stress, physical aggression, poor response to treatment, lengthy stay in the hospital, massive paranoia, little capacity for alcohol, no manic-depressive component, failure under adversity, discrepancy between ability and achievement, awareness of a change in the self, somatic delusions, a clash between the culture and the environment, and a loss of decency. In contrast, the reactive patient has these characteristics: good psychological history, good physical health, normal family member, well adjusted at school, domestic troubles unaccompanied by behavioral disruptions in the patient, extroverted behavior trends and interests, history of adequate social physical, and/or mental functioning, normal siblings, normally protective accepting mother, accepting father, heterosexual behavior, sudden onset of psychosis with pertinent stress present, verbal aggression, good response to treatment, short stay in the hospital, minor paranoid trends, good capacity for alcohol, manic-depressive component present, success despite adversity, harmony between ability and achievement, no sensation of self-change, absence of somatic delusions, harmony between the culture and the environment, and retention of decency.

The first three criteria apply to the patient's behavior between birth and the fifth year; the next seven, between the fifth year and adolescence; the next five, from adolescence to adulthood; the last nine, during adulthood. Using these 24 points to distinguish the two syndromes they tried to answer three questions:

1. Do diagnoses based upon the Rorschach alone label as nonpsychotic a portion of the population of mental patients who are clinically diagnosed as schizophrenic?

2. Can case histories of clinically diagnosed schizophrenics be differentiated into two categories: process and reactive?

3. Are those cases rated psychotic from the Rorschach classed as process on the basis of case histories, and are those cases judged nonpsychotic from the Rorschach classified as reactive from the case histories?

Two samples of 108 and 95 patients clinically diagnosed as schizophrenic were given the Rorschach and rated according to the process-reactive criteria. In the first sample of 108 patients, 57 were classified as psychotic and 51 nonpsychotic on the basis of the Rorschach alone, while in the second sample, of 74 patients who

could be rated as process or reactive, 36 were classified as psychotic, and 38 as nonpsychotic from their Rorschach protocols. Those patients who were rated as reactive from their history were most often judged nonpsychotic from the Rorschach, and those rated process from the case histories were most often judged as psychotic from the Rorschach.

Only one judge was used in the second sample to rate the patients as process or reactive, but two judges were used in the first sample. Of the 108 patients in this sample, both judges rated 86 cases, and were in agreement on 64 of these, which is greater than would be expected by chance.

However, the accuracy of the schizophrenic diagnosis is questionable in this study. If the Rorschach diagnosis is followed, then it appears that reactive schizophrenics are not psychotic. Furthermore, the psychiatric diagnosis appears to be somewhat contaminated because it was established on the basis of data collected by all appropriate services of the hospital, including psychological examinations. A similar type of contamination may have been present in classifying patients as process or reactive because one judge had reviewed each case previously and had seen psychological examination and history materials together prior to making his ratings. Three difficulties can be found with the criteria for process-reactive ratings. First, case histories are often incomplete and the patient is unable or unwilling to supply the necessary information. Second, it is difficult to precisely apply some of the criteria. For example, what is the precise dividing line between oddity and normality within the family? Third, in order to classify a patient it is necessary to set an arbitrary cut off point based on the number of process or reactive characteristics a patient has. Such a procedure needs validation.

Nonetheless, the results of this study support the view that schizophrenics can be classified as process or reactive, and that these syndromes differ in psychological functioning.

Another rating scale which has been used extensively to distinguish prognostically favorable and prognostically unfavorable schizophrenics was developed by Phillips (1953). The scale was developed from the case histories of schizophrenic patients who were eventually given shock treatment. The scale evaluates each patient in three areas: premorbid history, possible precipitating factors, and signs of the disorder. Premorbid history includes seven items on the social aspects of sexual life during adolescence and immediately beyond, seven items on the social aspects of recent sexual life, six items on personal relations, and six items on recent premorbid adjustment in personal relations. The sections of the scale which reflect the recent sexual life and its social history are the most successful in predicting the outcome of treatment. The items in the scales are arranged in order of increasing significance for improvement and nonimprovement away from the score of three, which is the dividing point between improved and unimproved groups. The premorbid history subscale has been utilized as the ranking instrument in the studies described by Rodnick and Garmezy (1957; Garmezy & Rodnick, 1959).

Another approach to the separation of schizophrenics into prognostic groups uses the activity of the autonomic nervous system as the basis for division (Meadow & Funkenstein, 1952; Meadow, Greenblatt, Funkenstein, & Solomon, 1953; Meadow,

Greenblatt, & Solomon, 1953). Meadow and Funkenstein (1952) worked with 58 schizophrenic patients tested for autonomic reactivity and for abstract thinking. Following therapy the patients were divided into two groups, good or poor, depending on the outcome of the treatment. The battery of psychological tests included the similarities and block design subtests of the Wechsler-Bellevue scale, the Benjamin Proverbs test, and the object sorting tests. The physiological test involved the systolic blood pressure reaction to adrenergic stimulation (intravenous Epinephrine) and cholinergic stimulation (intramuscular Mecholyl). On the basis of the physiological and psychological testing, schizophrenic cases were divided into three types: Type I, characterized by marked response to Epinephrine, low blood pressure, and failure of the blood pressure to rise under most stresses, loss of ability for abstract thinking, inappropriate affect, and a poor prognosis; Type II, characterized by an entirely different autonomic pattern, relatively intact abstract ability, anxiety or depression, and a good prognosis; Type III, showing no autonomic disturbance, relatively little loss of abstract ability, little anxiety, well organized paranoid delusions, and a fair prognosis.

However, as Meadow and Funkenstein (1952) point out, there is considerable overlap of the measures defining these types so that the classification must be tentative. Also, of the psychological tests used, only Proverbs distinguished significantly between the patients when they were classified according to autonomic reactivity, while Block Design failed to distinguish significantly among any of the types. Further research using this method of division (Meadow, Greenblatt, Funkenstein, & Solomon, 1953; Meadow, Greenblatt, & Solomon, 1953) served as a basis for investigations of the process-reactive syndromes by King (1958) and Zuckerman and Grosz (1959).

King (1958) hypothesized that predominantly reactive schizophrenics would exhibit a higher level of autonomic responsiveness after the injection of Mecholyl than predominantly process schizophrenics. The subjects were 60 schizophrenics who were classified as either process or reactive by the present investigator and an independent judge using the criteria of Kantor et al. (1953). Only those subjects were used on which there was classificatory agreement. This resulted in 22 process and 24 reactive patients. In order to consider the process-reactive syndrome as a continuum, 16 subjects were randomly selected from these two groups and were ranked by two independent raters.

While the patient was lying in bed shortly after awaking in the morning the resting systolic blood pressure was determined. The patient then received 10 milligrams of Mecholyl intramuscularly, and the systolic blood pressure was recorded at intervals up to 20 minutes. Then the maximum fall in systolic blood pressure (MFBP) below the resting blood pressure following the injection of Mecholyl was computed for the different time intervals. There was a significant difference in the MFBP score for the reactives as compared with the normals. For the 16 subjects, the correlation between the sets of ranks on the process-reactive dimension and MFBP was $-.58$.

In a second part of the study 90 schizophrenics, none of whom had participated in the first part, were classified as either process, process-reactive, or reactive, using the cri-

teria of Kantor et al. (1953). On this basis the subjects were divided into three groups of 24. Also, scores for 22 subjects were obtained on the Elgin Prognostic Scale, and 12 of these were rated independently by two raters. The MFBP scores were 17.04 for the process group, 22.79 for the process-reactive group, and 26.62 for the reactive. Using an analysis of variance a significant F score occurs at the .01 level. The correlation between the Elgin Prognostic Scale and the MFBP scores for 22 patients was −.49.

Results of both parts of the study revealed that the patients classified as reactive exhibited a significantly greater fall in blood pressure after the administration of Mecholyl than the process patients. This evidence points to diminished physiological responsiveness in process, but not in reactive schizophrenia. However, Zuckerman and Grosz (1959) found that process schizophrenics showed a significantly greater fall in blood pressure following the administration of Mecholyl than reactives. Since these results contradict King's findings the question of the direction of responsiveness to Mecholyl in these two groups requires further investigation before a conclusion can be reached.

Process-Organic Versus Reactive-Psychogenic

Brackbill and Fine (1956) suggested that process schizophrenics suffer from an organic impairment not present in the reactive case. They hypothesized that there would be no significant differences in the incidence of "organic signs" on the Rorschach between a group of process schizophrenics and a group of known cases of central nervous system pathology, and that both organic and process groups would show significantly more signs of organic involvement than the reactive group.

The subjects consisted of 36 patients diagnosed as process schizophrenics and 24 reactive schizophrenics. The criteria of Kantor et al. (1953) were used to describe the patients as process or reactive. Patients were included only when there was complete agreement between judges as to the category of schizophrenia. Also included in the sample were 28 cases of known organic involvement. All patients were given the Rorschach, and the protocols were scored using Piotrowski's (1940) 10 signs of organicity.

Using the criterion of five or more signs as a definite indication of organic involvement there was no significant difference between the organic and process groups, but both groups were significantly different from the reactives. Considering individual signs, four distinguished between the reactive and organic group, while two distinguished between process and reactive groups. The authors concluded that the results supported the hypothesis that process schizophrenics react to a perceptual task in a similar manner to that of patients with central nervous system pathology. No specific hypothesis was made about individual Rorschach signs, but color naming, completely absent in the reactives, was indicated as an example of concrete thinking and inability to abstract, suggesting that one of the critical differences between process and reactive groups is in terms of a type of thought disturbance.

This study does not provide detailed information about the manner of establishing the diagnosis of schizophrenia or about the judges deciding the process and reactive syndromes. Also, a further difficulty is the admitted inadequacy of the organic signs, since 66% of cases with organic

pathology in this study were false negatives according to the Rorschach criteria. Thus while the existence of the process and reactive syndromes is supported by the results of this investigation, there is less evidence of an organic deficit in process schizophrenics.

Becker (1956) pointed out that the consistency of the prognostic findings in schizophrenia has led to postulating two kinds of schizophrenia: process, with an organic basis, and reactive, with a psychological basis. He rejects this conclusion because research data in this area shows considerable group overlap, making it clinically difficult and arbitrary to force all schizophrenics into one group or the other. Also, if schizophrenia is a deficit reaction which may be brought about by any combination of 40 or more etiological factors, then the conception of two dichotomous types of schizophrenia is not useful. Finally, he maintains that 20 years of research have failed to find clear etiological differences between any subgroupings.

Instead, Becker stated that process and reactive syndromes should be conceived as end points on a continuum of levels of personality organization. Process reflects a very primitive undifferentiated personality structure, while reactive indicates a more highly organized one. He hypothesized that schizophrenics more nearly approximating the process syndrome would show more regressive and immature thinking processes than schizophrenics who more nearly approximate the reactive syndromes. His sample consisted of 51 schizophrenics, 24 males and 27 females, all under 41 years of age. Their thinking processes were evaluated by the Rorschach and the Benjamin Proverbs test. The 1937 Stanford-Binet vocabulary test was used to estimate verbal intelligence. A Rorschach scoring system was used which presumably reflected the subjects' level of perceptual development, while a scoring system was devised for the Proverbs which reflected levels of abstraction. Since there is a high relationship between intelligence and ability to interpret proverbs, a more sensitive index of a thinking disturbance was considered to be a discrepancy score based on the standard score difference between a vocabulary estimate of verbal intelligence and the proverbs score. Process and reactive ratings were made on the Elgin Prognostic Scale.

The Rorschach mean perceptual level score and the Elgin Prognostic Scale correlated −.599 for men and −.679 for women, indicating a significant relationship between the process-reactive dimension as evaluated from case history data and disturbances of thought processes as measured by the Rorschach scoring system. The proverbs-vocabulary discrepancy score was significantly related to the process-reactive dimension for men, but not for women. No adequate explanation was found for this sex difference, which mitigates the results. A further difficulty occurs because the case history and test evaluations were made by the same person. However, the results in part support the hypothesis, indicating evidence for a measurable dimension of regressive and immature thinking related to the process-reactive dimension.

McDonough (1960), acting on the assumption that process schizophrenia involves central nervous system pathology specifically cortical in nature, hypothesized that brain damaged patients and process schizophrenics would have significantly lower critical flicker frequency (CFF) thresholds and would be unable to

perceive the spiral aftereffect significantly more often than reactive schizophrenics and normals. Four groups of 20 subjects each were tested. The organic group consisted of individuals with known brain damage. One hundred and sixty-one schizophrenic case histories were examined, and 76 were chosen from this group to be rated on the Elgin Prognostic Scale. The 20 patients receiving the lowest point totals were selected as being most reactive, while those with the 20 highest scores were considered most process.

Results of the experiment revealed that organic patients were significantly different from all other groups in CFF threshold and ability to perceive the spiral aftereffect. Process and reactive schizophrenics did not differ from each other on either task, but reactive schizophrenics had higher CFF thresholds than normals. These results do not indicate demonstrable cortical defect in either process or reactive schizophrenia.

Process-Poor Premorbid History Versus Reactive-Good Premorbid History

Rodnick and Garmezy (1957), discussing the problem of motivation in schizophrenia, reviewed a number of studies in which the Phillips prognostic scale was used to classify schizophrenic patients into two groups, good and poor. For example, Bleke (1955) hypothesized that patients whose prepsychotic life adjustment was markedly inadequate would have greater interferences and so show more reminiscence following censure than patients whose premorbid histories were more adequate.

The subjects were presented with a list of 14 neutrally toned nouns projected successively on a screen. Each subject was required to learn to these words a pattern of pull-push movements of a switch lever. For half the subjects in each group learning took place under a punishment condition, while the remaining subjects were tested under a reward condition. The subjects consisted of 40 normals, 20 poor premorbid schizophrenics, and 20 good premorbid schizophrenics. The results confirmed the hypothesis.

A reanalysis of Dunn's (1954) data indicated that a poor premorbid group showed discrimination deficits when confronted with a scene depicting a mother and a young boy being scolded, but good premorbid and normal subjects did not show this deficit.

Mallet (1956) found that poor premorbid subjects in a memory task for verbal materials showed significantly poorer retention of hostile and nonhostile thematic contents than did good premorbid and normal subjects. Harris (1955) has found that in contrast to goods and normals poor premorbids have more highly deviant maternal attitudes. They attribute more rejective attitudes to their mothers, and are less able to critically evaluate their mothers. Harris (1957) also found differences among the groups in the size estimation of mother-child pictures. The poors significantly overestimated, while the goods underestimated, and the normals made no size error.

Rodnick and Garmezy (1957) reported a study using Osgood's (1952) semantic differential techniques in which six goods and six poors rated 20 concepts on each of nine scales selected on the basis of high loadings on the evaluative, potency, and activity factors. Good and poor groups differed primarily on potency and activity factors. The poors described words with negative value, as more powerful and active. The goods could discriminate among concepts, but the

poors tended to see most concepts as powerful and active.

Rodnick and Garmezy (1957) also investigated differences in authority roles in the family during adolescence in good and poor premorbid patients. While results were tentative at that time, they suggested that the mothers of poor premorbid patients were perceived as having been more dominating, restrictive, and powerful, while the fathers appeared ineffectual. The pattern was reversed in the good premorbid patients.

Alvarez (1957) found significantly greater preference decrements to censured stimuli by poor premorbid patients. This result was consistent with the results of Bleke's (1955) and Zahn's (1959) observations of reversal patterns of movement of a switch lever following censure. These experiments suggested an increased sensitivity of the poor premorbid schizophrenic patient to a threatening environment.

These studies reported by Rodnick and Garmezy (1957) indicated that it was possible, using the Phillips scale, to effectively dichotomize schizophrenic patients. However, the Phillips scale had predictive validity only when applied to male patients. Within this form of reference it was also possible to demonstrate differences between goods and poors in response to censure, and in perception of familial figures. Variability in the results of schizophrenic performance was considerably reduced by dichotomizing the patients, but it was often impossible to detect significant differences between the performance of good premorbid schizophrenics and normals. Rodnick and Garmezy (1957) suggest that the results be considered as preliminary findings pending further corroboration, though providing support for the concept of premorbid groups of schizophrenics differing in certain psychological dimensions.

Process-Reactive Empirical-Theoretical Formulations

Fine and Zimet (1959; Zimet & Fine, 1959) used the same population employed by Kantor et al. (1953) and the same criteria for distinguishing the process and reactive patients. For this study only those cases were included where there was complete agreement among the judges as to the category of schizophrenia. They studied the level of perceptual organization of the patients as shown on their Rorschach records. The process group was found to have significantly more immature, regressive perceptions, while the reactive group gave more mature and more highly organized responses. The findings indicated that archaic and impulse-ridden materials break through more freely in process schizophrenia, and that there is less ego control over the production of more regressive fantasies. Zimet and Fine (1959) speculated that process schizophrenia mirrors oral deprivation of early ego impoverishment, so that either regression or fixation to an earlier developmental stage is reflected in his perceptual organization. In contrast, it is possible that the reactive schizophrenic's ego weakness occurs at a later stage in psychosexual development, and any one event may reactivate the early conflict.

An amplification of the process-reactive formation has been suggested by Kantor and Winder (1959). They hypothesized that schizophrenia can be understood as a series of responses reflecting the stage of development in the patient's life at which emotional support was severely deficient. Schizophrenia can be quantitatively depicted in terms of the level in life to which the schizophrenic has regressed,

and beyond which development was severely distorted because of disturbing life circumstances. The earlier in developmental history that severe stress occurs, the more damaging the effect on subsequent interpersonal relationships. Sullivan (1947) suggested five stages in the development of social maturity: empathic, prototaxic, parataxic, autistic, and syntaxic. The most malignant schizophrenics are those who were severely traumatized in the empathic stage of development when all experience is unconnected, there is no symbolism, and functioning is at an elementary biological level. The schizophrenic personality originating at this stage may show many signs of organic dysfunction. Prognosis will be most unfavorable, and delusional formation will tend to be profound.

In view of the primitive symbolic conduct and the lack of a self-concept in the prototaxic stage, the schizophrenic personality referable to this stage will be characterized by magical thinking and disturbed communication. The delusion of adoption often occurs. However, these patients are more coherent than those of the previous level.

The parataxic schizophrenic state involves the inability of the self-system to prevent dissociation. The autonomy of the dissociations result in the patient's fear of uncontrollable inward processes. Schizophrenic symptoms appear as regressive behavior attempting to protect the self and regain security in a threatening world. Delusional content usually involves world disaster coupled with bowel changes. Nihilistic delusions are common. While there is evidence of a self-system in these patients, prognosis remains unfavorable.

The patient who has regressed to the autistic stage, although more reality oriented than in the previous stages, is characterized by paranoid suspiciousness, hostility, and pathological defensiveness against inadequacy feelings. A consistent system of delusions will be articulated and may bring the patient into conflict with society. However, prognosis is more favorable at this stage than previously.

An individual at the syntaxic level has reached concensus with society, so that if schizophrenia occurs it will be a relatively circumscribed reaction. Onset will be sudden with plausible environmental stresses, and prognosis is relatively good.

Becker (1959) also elaborated on the lack of a dichotomy in schizophrenia. Individual cases spread out in such a way that the process syndrome moves into the reactive syndrome, so that the syndromes probably identify the end points of a dimension of severity. At the process end of the continuum the development of personality organization is very primitive, or involves severe regression. There is a narrowing of interests, rigidity of structure, and inability to establish normal heterosexual relationships and independence. In contrast, the reactive end of the continuum represents a higher level of personality differentiation. The prepsychotic personality is more normal, heterosexual relations are better established, and there is greater tolerance of environmental stresses. The remains of a higher developmental level are present in regression and provide strength for recovery.

Becker (1959) factor analyzed some of the data from his previous study (Becker, 1956). The factored matrix included a number of background variables, the 20 Elgin Prognostic Scale subscores, and a Rorschach genetic level score (*GL*) based on the first response to each card. Seven centroid factors were extracted

from the correlation matrix. Factors 4, 6, and 7 represented intelligence, cooperativeness, and marital status of parents, respectively. The highest loadings on Factor 5 were history of mental illness in the family, excellent health history, lack of precipitating factors, and clouded sensorium. The Rorschach *GL* score and the Elgin scales did not load significantly on Factors 4 through 7.

The remaining three factors parallel the factors Lorr, Wittman, and Schanberger (1951) found with 17 of the 20 Elgin scales using an oblique solution instead of the orthogonal solution used in this study. Factor 1 is called schizophrenic withdrawal, loading on defect of interest, insidious onset, shut-in personality, long duration of psychosis, and lack of precipitating factors. At one end this factor defines the typical process syndrome, while the other end describes the typical reactive syndrome. The Rorschach *GL* score loaded −.46 on Factor 1.

Factor 2, reality distortion, loads on hebephrenic symptoms, bizarre delusions, and inadequate affect. Rorschach *GL* score loaded −.64 on this factor. Factor 3 loaded on indifference and exclusiveness-stubbornness. The opposite pole of this factor involves insecurity, inferiority, self-consciousness, and anxiety. Rorschach *GL* score loaded .25 on this factor.

Further analysis indicated that when Factors 1 and 2 were plotted against each other an oblique rotation was required, introducing a correlation of from .60 to .70 between schizophrenic withdrawal and reality distortion factors. Similar obliqueness was found between Factors 2 and 3, suggesting the presence of a second-order factor.

However, the sampling of behavior manifestations in the Elgin scale overweights the withdrawal factor, which gives Factor 1 undue weight and biases the direction of a second-order factor toward the withdrawal factor. Also, it is not possible to accurately locate second-order factors with only seven first-order factors as reference points. In addition, sample size and related sampling errors limited inferences about a second-order factor. There is the suggestion, however, of the existence of a general severity factor, loading primarily schizophrenic withdrawal and reality distortion.

The author suggests utilizing the evidence from this study to form an index of severity of psychosis which could be used to make diagnoses with prognostic significance. This diagnostic procedure would include factor estimates of schizophrenic withdrawal and emotional rigidity, based on Elgin scale ratings, and reality distortion, based on the Rorschach *GL* score.

Garmezy and Rodnick (1959) pointed out that despite failure to find support for a fundamental biological deviation associated with schizophrenia (Kety, 1959), the view of schizophrenia as a dichotomous typology influenced either by somatic or psychic factors has continuously been advanced. They maintain that on the basis of empirical evidence there is little support for a process-organic versus reactive-psychogenic formulation of schizophrenic etiology.

Reviewing a series of studies using the Phillips scale as a dichotomizing instrument (Alvarez, 1957; Bleke, 1955; Dunham, 1959; Dunn, 1954; Englehart, 1959; Farina, 1960; Garmezy, Stockner, & Clarke, 1959; Harris, 1957; Kreinik, 1959; Rodnick & Garmezy, 1957; Zahn, 1959) Garmezy and Rodnick concluded that the results indicate two groups of schizophrenic patients differing both

in prognostic potential and sensitivity to experimental cues. There is an interrelationship among the variables of premorbid adequacy, differential sensitivity to censure, prognosis, and types of familial organization. This suggests a relationship between varying patterns of early experience and schizophrenia, though it does not embody the acceptance of a given position regarding psychological or biological antecedents in schizophrenia.

Reisman (1960), in an attempt to explain the heterogeneous results of psychomotor performance in schizophrenics, suggested that there were two groups of schizophrenics, process and reactive, differing in motivation. The process group was seen as more withdrawn and indifferent to their performance, and consequently reflecting a psychomotor deficit not present in reactives. In order to test this hypothesis 36 reactives, 36 process patients, and 36 normals performed a card-sorting task. The groups were distinguished according to the criteria of Kantor, Wallner, and Winder (1953). On Trial 1 all subjects were requested to sort as rapidly as possible. Then the subjects were assigned to one of four experimental conditions, with an attempt made to equate across the experimental conditions for age, estimated IQ, length of hospitalization, and initial sorting time. Condition 1 (FP) involved sorting the cards seven more times and if the sort was fast the subjects were shown stress-arousing photographs. If they sorted slowly no photographs were shown. Condition 2 (SP) was the reverse of this. Condition 3 (FL) and Condition 4 (SL) were similar to the first two conditions except that a nonreinforcing light was used instead of the pictures. After Trial 8 all subjects were informed that there would be no more pictures or light, but were asked to sort rapidly for three more trials. With four conditions on Trials 2 through 8, 10 subjects from each of the three groups participated in each of the two picture conditions, while eight subjects from each group participated in each of the light conditions.

The results indicated that the normals performed about the same under all conditions. The process group under FP sorted as fast as normals, but performed slowly under the other three conditions, while the reactives were slowest under FP but were as fast as normals under the other three conditions. Within all three groups performance under FL did not differ significantly from performance under SL. Under FL and SL, however, reactives and normals sorted more rapidly than the process group. These results supported the hypothesis of a motivational deficit for process schizophrenics. The results also indicated that the pictures were negatively reinforcing for the reactives, while the process patients were motivated to see them. This suggested a withdrawal differential. The withdrawal of the process patients is of such duration that supposedly threatening photographs cause little anxiety. In contrast, reactive withdrawal is motivated by an environment that recently became unbearable. Confronted with pictures representing this environment the reactive patient experiences anxiety and avoidance. However, the results of this experiment are in contrast to the findings of Rodnick and Garmezy (1957) that prolonged exposure to social censure will result in greater sensitivity to that stimulation.

Summary

This review of all the research on the process-reactive classification of

schizophrenia strongly indicates that it is possible to divide schizophrenic patients into two groups differing in prognostic and life-history variables. Using such a division it is also possible to demonstrate differences between the two groups in physiological measures and psychological dimensions.

The result of such an approach has been to clarify many of the heterogenous reactions found in schizophrenia. It also appears that the dichotomy is somewhat artificial and really represents end points on a continuum of personality organization. The most process patient represents the extreme form of personality disintegration, while the most reactive patient represents the extreme form of schizophrenic integration. The reactions of this type of patient are often difficult to distinguish from behavior patterns of normal subjects. There does not appear to be any significant evidence to support the contention of a process-organic versus a reactive-psychogenic formulation of schizophrenic etiology.

It is difficult to decide on the most appropriate criteria for selecting schizophrenic subjects so as to reduce their response variability. Preferences are generally found for one of three sets of criteria: Kantor, Wallner, and Winder's (1953) items, the Elgin Prognostic Scale (1944), or the Phillips scale (1953). The criteria of Kantor et al. (1953) does not provide a quantitative ordering of the variables, and is descriptively vague in several dimensions as well as depending upon life history material which is not always available. While the Elgin scale does provide a quantitative approach, it also has the disadvantages of descriptive vagueness and excessive dependence upon life history material. The Phillips scale eliminates some of these difficulties, but its validity is limited to the adequacy or inadequacy of social-sexual premorbid adjustment. The need for more feasible criteria may be met by the factor analysis of pertinent variables to obtain a meaningful severity index (Becker, 1959), or by using rating scales in which the patient verbally supplies the necessary information. An example of the latter is the Ego Strength scale (Barron, 1953), recently utilized in distinguishing two polar constellations of schizophrenia; a process type with poor prognosis and grossly impaired abstract ability, and a reactive type characterized by good prognosis and slight abstractive impairment (Herron, in press).

This need for more efficient differentiating criteria mitigates some of the significance of present findings using the process-reactive dimension. Nonetheless, the process-reactive research up to this time has succeeded in explaining schizophrenic heterogeneity in a more meaningful manner than previous interpretations adhering to various symptom pictures and diagnostic subtypes. Consequently, there appears to be definite value in utilizing the process-reactive classification of schizophrenia.

REFERENCES

Alvarez, R. R. A comparison of the preferences of schizophrenic and normal subjects for rewarded and punished stimuli. Unpublished doctoral dissertation, Duke University, 1957.

Barron, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, **17**, 327–333.

Becker, W. A genetic appoach to the interpretation and evaluation of the process-reactive distinction in schizophrenia. *J. abnorm. soc. Psychol.*, 1956, **53**, 229–236.

BECKER, W. C. The process-reactive distinction: A key to the problem of schizophrenia? *J. nerv. ment. Dis.*, 1959, **129**, 442–449.

BENJAMIN, J. D. A method for distinguishing and evaluating formal thinking disorders in schizophrenia. In J. S. Kasanin (Ed.), *Language and thought in schizophrenia.* Berkeley: Univer. California Press, 1946. Pp. 66–71.

BLEKE, R. C. Reward and punishment as determiners of reminiscence effects in schizophrenics and normal subjects. *J. Pers.*, 1955, **23**, 479–498.

BLEULER, E. *Dementia praecox.* New York: International Univer. Press, 1911.

BRACKBILL, G., & FINE, H. Schizophrenia and central nervous system pathology. *J. abnorm. soc. Psychol.*, 1956, **52**, 310–313.

CHASE, L. S., & SILVERMAN, S. Prognosis in schizophrenia: An analysis of prognostic criteria in 150 schizophrenics treated with Metrazol or insulin. *J. nerv. ment. Dis.*, 1943, **98**, 464–473.

DUNHAM, R. M. Sensitivity of schizophrenics to parental censure. Unpublished doctoral dissertation, Duke University, 1959.

DUNN, W. L. Visual discrimination of schizophrenic subjects as a function of stimulus meaning. *J. Pers.*, 1954, **23**, 48–64.

ENGLEHART, R. S. Semantic correlates of interpersonal concepts and parental attributes in schizophrenia. Unpublished doctoral dissertation, Duke University, 1959.

FARINA, A. Patterns of role dominance and conflict in parents of schizophrenic patients. *J. abnorm. soc. Psychol.*, 1960, **61**, 31–38.

FINE, H. J., & ZIMET, C. N. Process-reactive schizophrenia and genetic levels of perception. *J. abnorm. soc. Psychol.*, 1959, **59**, 83–86.

GARMEZY, N., & RODNICK, E. H. Premorbid adjustment and performance in schizophrenia: Implications for interpreting heterogeneity in schizophrenia. *J. nerv. ment. Dis.*, 1959, **129**, 450–466.

GARMEZY, N., STOCKNER, C., & CLARKE, A. R. Child-rearing attitudes of mothers and fathers as reported by schizophrenic and normal control patients. *Amer. Psychologist*, 1959, **14**, 333. (Abstract)

HARRIS, J. G., JR. A study of the mother-son relationship in schizophrenia. Unpublished doctoral dissertation, Duke University, 1955.

HARRIS, J. G., JR. Size estimation of pictures as a function of thematic content for schizophrenic and normal subjects. *J. Pers.*, 1957, **25**, 651–672.

HERRON, W. G. Abstract ability in the process-reactive classification of schizophrenia. *J. gen. Psychol.*, in press.

HUNT, R. C., & APPEL, K. E. Prognosis in psychoses lying midway between schizophrenia and manic-depressive psychoses. *Amer. J. Psychiat.*, 1936, **93**, 313–339.

KANT, O. Differential diagnosis of schizophrenia in light of concepts of personality stratification. *Amer. J. Psychiat.*, 1940, **97**, 342–357.

KANT, O. A comparative study of recovered and deteriorated schizophrenic patients. *J. nerv. ment. Dis.*, 1941, **93**, 616–624.

KANT, O. The evaluation of prognostic criteria in schizophrenia. *J. nerv. ment Dis.*, 1944, **100**, 598–605.

KANTOR, R., WALLNER, J., & WINDER C. Process and reactive schizophrenia. *J. consult. Psychol.*, 1953, **17**, 157–162.

KANTOR, R. E., & WINDER, C. L. The process-reactive continuum: A theoretical proposal. *J. nerv. ment. Dis.*, 1959, **129**, 429–434.

KETY, S. S. Biochemical theories of schizophrenia. *Science*, 1959, **129**, 1528–1532, 1590–1596, 3362–3363.

KING, G. Differential autonomic responsiveness in the process-reactive classification of schizophrenia. *J. abnorm. soc. Psychol.*, 1958, **56**, 160–164.

KING, G. F. Research with neuropsychiatric samples. *J. Psychol.*, 1954, **38**, 383–387.

KREINIK, P. S. Parent-child themas and concept attainment in schizophrenia. Unpublished doctoral dissertation, Duke University, 1959.

KRETSCHMER, E. *Physique and character.* New York: Harcourt, Brace, 1925.

LANGFELDT, G. The diagnosis of schizophrenia. *Amer. J. Psychiat.*, 1951, **108**, 123–125.

LEWIS, N. D. C. *Research in dementia praecox.* New York: Natronal Committee for Mental Hygiene, 1936.

LEWIS, N. D. C. The prognostic significance of certain factors in schizophrenia. *J. nerv. ment. Dis.*, 1944, **100**, 414–419.

LORR, M., WITTMAN, P., & SCHANBERGER, W. An analysis of the Elgin Prognostic scale. *J. clin. Psychol.*, 1951, **7**, 260–263.

MCDONOUGH, J. M. Critical flicker frequency and the spiral aftereffect with process and reactive schizophrenics. *J. consult. Psychol.*, 1960, **24**, 150–155.

MALAMUD, W., & RENDER, N. Course and prognosis in schizophrenia. *Amer. J. Psychiat.*, 1939, **95**, 1039–1057.

MALLET, J. J. Verbal recall of hostile and neutral thematic contents by schizophrenic and normal subjects. Unpublished doctoral dissertation, Duke University, 1956.

MAUZ, F. *Die Prognostik der endogen Psychosen.* Leipzig: G. Theime, 1930.

MEADOW, A., & FUNKENSTEIN, D. H. The

relationship of abstract thinking to the automatic nervous system in schizophrenia. In P. H. Hoch and J. Zubin (Eds.), *Relation of psychological tests to psychiatry.* New York: Grune & Stratton, 1952. Pp. 131–144.

MEADOW, A., GREENBLATT, M., FUNKENSTEIN, G. H., & SOLOMON, H. C. The relationship between the capacity for abstraction is schizophrenia and the physiologic response to autonomic drugs. *J. nerv. ment. Dis.*, 1953, **118**, 332–338.

MEADOW, A., GREENBLATT, M., & SOLOMON, H. C. "Looseness of association" and impairment in abstraction in schizophrenia. *J. nerv. ment. Dis.*, 1953, **118**, 27–35.

MILICI, P. Postemotive schizophrenia. *Psychiat. Quart.*, 1939, **13**, 278–293.

OSGOOD, C. E. The nature and measurement of meaning. *Psychol. Bull.*, 1952, **49**, 197–237.

PASKIND, J. A., & BROWN, M. Psychosis resembling schizophrenia occurring with emotional stress and ending in recovery. *Amer. J. Psychiat.*, 1940, **96**, 1379–1388.

PHILLIPS. L. Case history data and prognosis in schizophrenia. *J. nerv. ment. Dis.*, 1953, **117**, 515–525.

PIOTROWSKI, Z. A. Positive and negative Rorschach organic reactions. *Rorschach res. Exch.*, 1940, **4**, 143–151.

RABIN, A. J., & KING, G. F. Psychological studies. In L. Bellak (Ed.), *Schizophrenia: A review of the syndrome.* New York: Logos, 1958. Pp. 216–278.

REISMAN, J. M. Motivational differences between process and reactive schizophrenics. *J. Pers.*, 1960, **28**, 12–25.

RODNICK E. H., & GARMEZY, N. An experimental approach to the study of motivation in schizophrenia. In M. R. Jones (Ed.), *Nebraska symposium on motivation: 1957.* Vol. V. Lincoln: Univer. Nebraska Press, 1957. Pp. 109–184.

SULLIVAN, H. S. *Conceptions of modern psychiatry.* Washington: W. A. White Psychiatric Foundation, 1947.

WITTMAN, P. A scale for measuring prognosis in schizophrenic patients. *Elgin State Hosp. Pap.*, 1941, **4**, 20–33.

WITTMAN, P. Follow-up on Elgin prognosis scale results. *Ill. psychiat. J.*, 1944, **4**, 56–59.

WITTMAN, P., & STEINBERG, L. Follow-up of an objective evaluation of prognosis in dementia praecox and manic-depressive psychoses. *Elgin State Hosp. Pap.*, 1944, **5**, 216–227. (a)

WITTMAN, P., & STEINBERG, D. L. Study of prodromal factors in mental illness with special references in schizophrenia. *Amer. J. Psychiat.*, 1944, **100**, 811–816. (b)

ZAHN, T. P. Acquired and symbolic affective value as determinant of size estimation in schizophrenic and normal subjects. *J. abnorm. soc. Psychol.*, 1959, **58**, 39–47.

ZIMET, C. N., & FINE, H. J. Perceptual differentiation and two dimensions of schizophrenia. *J. nerv. ment. Dis.*, 1959, **129**, 435–441.

ZUCKERMAN, M., & GROSZ, H. J. Contradictory results using the mecholyl test to differentiate process and reactive schizophrenia. *J. abnorm. soc. Psychol.*, 1959, **59**, 145–146.

(Received March 29, 1961)

Psychological Bulletin
1962, Vol. 59, No. 4, 344–352

A PARADIGM FOR DETERMINING THE CLINICAL RELEVANCE OF HYPNOTICALLY INDUCED PSYCHOPATHOLOGY

JOSEPH REYHER
Michigan State University

Hypnotic research can be broadly characterized as having either an intrinsic or instrumental orientation. Intrinsically oriented research is concerned with the phenomena and nature of hypnosis itself whereas instrumentally oriented research utilizes hypnosis to produce some condition which is the object of study, e.g., personality alteration, psychopathology. Although both research orientations present difficult methodological problems, this communication will limit itself to problems associated with the instrumental use of hypnosis.

Despite its ability to command enduring interest, instrumental hypnotic research has remained relatively inconsequential and isolated. One of the principle reasons for this state of affairs is the lack of criteria for determining the relevance of hypnotically induced behavior to clinical or natural behavior. In the absence of adequate criteria, the data of instrumental hypnotic research tend to be either rejected or consigned to the limbo of ambiguity. Adams (1957), in his review of laboratory studies of behavior without awareness, excluded studies involving posthypnotic suggestion, automatic writing, extrasensory perception, and processes of which the subject is unaware. A paralyzing caution was displayed by Ainsworth (1954) in her review of Rorschach validation research:

> Hypnosis provides another method for artifically altering the state of the subject while undergoing the Rorschach examination, although hypnotic studies are open to the question of whether the hypnotically induced state is comparable enough to the "genuine" state to provide validation evidence (p. 480).

Most reviewers, however, do not even bother to mention the exclusion of hypnotic research. At the risk of being charitable, it is likely that rejecting attitudes toward instrumental hypnotic research arise more from the lack of criteria for determining relevance than from prejudice. In lieu of such criteria, most investigators have approached this issue by ignoring it or by assuming that the induced behavior is equivalent in all respects to its natural counterpart. Phenotypic identity, however, does not necessarily imply genotypic identity; i.e., the fact that behavior similar to anxiety can be produced by hyposis does not mean that the mechanisms of hypnosis are the same or similar to the processes underlying clinical anxiety. Weitzenhoffer (1953) has pointed out that hypnotically induced phenomena resembling psychodynamic manifestations are apt to lack affective tone. He also recognized the importance of inducing an appropriate genotype by his assertion that affective tone is most apt to be absent "when the suggestions are aimed at directly bringing about the overt manifestations rather than creating the type of factors normally responsible for these" (p. 217).

The topic of hypnotically induced psychopathology will serve as the focus of the inquiry because it highlights both the methodological and conceptual problems involved in the laboratory investigation of hypnot-

ically induced conditions. Such a focus achieves enhanced significance because the genotypic-phenotypic relationships that constitute psychopathology represent one of the central problems in most psychoanalytically oriented theories of personality.

A Paradigm for the Hypnotic Induction of Psychopathology

A paradigm for demonstrating valid psychopathology must include a procedure for separating the mechanisms of suggestion from the mechanisms of pathogenic psychodynamics. Although it is doubtful that the mechanisms of hypnotic suggestion are similar to the mechanisms of pathogenic psychodynamics, clinical experience with hypnosis (Eisenbud, 1937; Rosen, 1953) indicates that hypnotic suggestion can set in motion nonsuggested pathogenic psychodynamics and observable psychopathology. Thus, hypnotic suggestion should be used only to *induce* a process that, under certain specifiable conditions, is theoretically capable of *producing* pathogenic psychodynamics and psychopathology. The hypnotically induced process defines the genotype, and the behavioral outcome defines the phenotype. The genotype is defined operationally by the statements in the hypnotic suggestions; the phenotype is defined operationally by a description of the subject's overt behavior. The description of the phenotype is considered to be operationally valid clinical psychopathology only if it satisfies the defining criteria of a given classification of psychopathology. In this way, the investigator can operationally tie down the genotype, or psychodynamics, that produces the observed psychopathology instead of having to relie upon the uncertainties of clinical inference in regard to natural psychopathology.

The production of operationally valid clinical psychopathology by a hypnotically induced process permits the inference that the genotype is adequate, which in turn is supporting evidence for the theory from which the genotype is derived. If the genotype does not produce psychopathology, there are two interpretive alternatives available: the genotype is inadequate and the theory from which it is derived is not supported, or the conditions of the experiment were unfavorable for an adequate test of the theory.

The foregoing considerations suggest four principles, or criteria, that should guide research in this area. First, the induced process must in no way include cues as to how the experimenter expects the subject to respond in any other respect. Orne (1959) has demonstrated convincingly the sensitivity of hypnotized subjects to the expectations of the experimenter and the "demand" characteristics of the experimental design. Second, the induced process must produce other processes and behavior; that is, it must be response-producing. Third, some of these responses must satisfy the defining criteria for inclusion in some classification of psychopathology. Finally, as Orne (1959) suggests, some of the subjects must be asked by a co-experimenter, unknown to the experimenter, to fake hypnosis in order to determine the demand characteristics of the research.

Review of Relevant Research

Research in the area of hypnotically induced psychopathology falls into three categories.

Direct Suggestion

In studies of this type, suggestion is used to produce a given response

which is considered to be clinically meaningful. By suggestion the experimenter reproduces in the subject a specific mood, attitude, affect, or symptom. Although most of this research has been reviewed elsewhere (Weitzenhoffer, 1953), a recent investigation by Levitt, den Breeijen, and Persky (1960) will be presented and discussed in detail because it is a particularly good example of the inherent defects in this popular approach. The procedure is straightforward: the subject is made to feel anxious by listening to a taped presentation of a series of somewhat repetitious phrases of increasing emotional intensity containing a variety of synonyms for the emotions of anxiety and fear.

A deliberate effort was made to produce anxiety in "pure" form because, under natural conditions, there is usually an admixture of anxiety, depression, hostility, etc. Their attempt to produce anxiety in pure form is, therefore, an interference with the idiosyncratic phenotype and entirely destroys its clinical significance. Due to its covert intrapsychic origins, anxiety is not experienced in the same way by everyone, nor is its presence always detected and identified. Moreover, their emphasis upon such words as "fear," "dread," "apprehension," and "panic" may well be reproducing responses to external threat rather than generating responses to an unknown internal threat, which is a distinction often used to differentiate anxiety from fear. The affect of anxiety is even more complicated than they have observed because it can be managed in different ways. It may be managed defensively by hostile or depressive reactions, projected as in a phobia, or converted into somatic processes. A study which ignores the personal equation in the hypnotic production of psychopathology should be designated as an experimental analogue of the clinical behavior in question. It is scientifically legitimate to endeavor to abstract and to purify emotions as they have done, but these, by definition, are not clinical phenomena; it is controlled, laboratory research which purposely creates conditions to eliminate the clinical "taint" of the data.

In terms of the paradigm, the major shortcoming of direct suggestion is the identity between the genotype and phenotype. The subject merely carries out the suggestions that are given to him; the instructions specify the behavior. This also means that direct suggestion is not response-producing in the sense that other processes are set in motion which lead to the behavioral outcome. In order to satisfy the paradigm, a process must be induced that has the capacity to trigger off a chain of events that eventuates in psychopathology. The nature of the genotype and the conditions under which it is induced will reflect some theory about personality and psychopathology. In this sense, the paradigm is a procedure for testing theories. Direct suggestion tests nothing but itself.

The Induction of Artificial Conflicts

In studies of this type (Bobbitt, 1958; Counts & Mensh, 1950; Erickson, 1944; Huston, Shakow, & Erickson, 1934; Luria, 1932), the subject is provided with a paramnesia regarding a situation to which he has a distressing emotional reaction, such as hostility or remorse. In one way or another, the subject is usually told that he will not remember anything about the experience posthypnotically, but, nevertheless, it will be disturbing to him. Although the induced experiences are intended to be perceived as "real" rather than con-

trived, the paradigm is not satisfied: the subject is told that he will not recall the paramnesia or that he will recall it to a certain degree; furthermore, he is told that the paramnesia will be a source of posthypnotic disturbance. The design of these studies is also incomplete because of the lack of control subjects who are asked to fake hypnosis, and the significance of the results is vitiated by the relatively weak disturbances that were produced.

Wohlberg (1947) reported a procedure which seems to approach closely the paradigm. Instead of implanting a paramnesia, he suggested an impulse that would produce conflict in the waking state. His instructions were as follows:

> When you awaken you will find next to you a bar of chocolate. You will have a desire to eat the chocolate that will be so intense that it will be impossible to resist the craving. At the same time you will feel that the chocolate does not belong to you and that to eat it would be very wrong and very bad. You will have no memory of these suggestions when you awaken, but you will, nevertheless, react to them (p. 337).

The distinctive aspect of his instructions is the posthypnotic suggestion of an overwhelming impulse which is rendered anxiety-producing by pitting it against conscience.

Although his subjects were instructed to perceive the induced impulse in terms of conscience, they were not instructed to develop symptoms. Accordingly, it is of great interest that the procedure spontaneously produced both somatic and psychological reactions, which included such marked symptoms as dizziness, tachycardia, and a negative hallucination. Since his procedure approximates closely the paradigm, the posthypnotic psychopathology may very well be a valid clinical phenomenon. If he had used the proper control subjects, a more positive statement could be made. In order for his instructions to be perfect, the subjects should not have been told how to perceive the impulse nor should he have suggested an amnesia. The impulse should spontaneously generate all subjects' reactions.

An investigation by Reyher (1961) also approaches the paradigm. Under deep hypnosis the subjects were given a hallucinatory experience that generated intense feelings of hositility toward a given individual. The instructions were as follows:

> Now listen carefully. After I awaken you, you will not be able to remember anything about this session. However, anything that comes into your conscious mind that is associated with this experience [specific classes of words are mentioned] will stir up overwhelming feelings of hate. If these feelings break into consciousness, you will realize that it is the person who owns these papers (which are within arm's reach] that you hate, and you will have an overwhelming urge to tear them up.

Posthypnotic conflict was created by presenting trachistoscopically critical and neutral pairs of words until one word of each pair was recognized. Ideally, the instructions should not have included such an ambiguous word as "if," nor should an amnesia have been produced, even though the conflict-producing impulse was to be experienced and acted upon psothypnotically. Nevertheless, the procedure produced much psychopathology. The recognition of conflict words produced such reactions as urticaria, tachycardia, gastric distress, headache, flushing, sweating, tics, tremors, and such psychological reactions as anxiety, apprehension, dissociation, and derivatives. One of the most important findings was a correlation of .74 between the degree repression of the induced conflict and the proportion of somatic complaints.

Other than the use of the ambiguous word if in the hypnotic instruc-

tions and the suggested amnesia, this study satisfies the paradigm. The induced hostility contained no clues in relation to the occurrences of psychopathology, many symptoms were produced and proper control subjects did not report symptoms. Accordingly, it is reasonable to conclude that the psychopathology was genuine. Spontaneous symptomatic reactions to hypnotically induced processes have been summarized by Weitzenhoffer (1953). Although these case reports of idiosyncratic reactions to hypnotic procedures are illuminating, they do not lend themselves to laboratory investigation because of their unreliable and uncontrolled nature.

The Activation of Natural Conflicts

Two studies fall into this category. Gordon (1959) instructed his subjects to bring to mind episodes involving conflict with parents. The subjects were given differing degrees of posthypnotic awareness of the episodes. No symptoms were reported. Although this approach gains clinical significance by permitting the subject to dwell upon his own emotionalized experiences, the investigation is difficult to interpet because of the posthypnotic suggestion to achieve a given degree of awareness. Utilizing the subject's own conflicts is a promising method for producing psychopathology and eliminates almost entirely the criticism of artificiality, provided that the subject is not instructed how to react posthypnotically. Nevertheless, it must be shown that such reliving of past experiences is anxiety-producing. Since no symptoms were reported, the conflicts were probably not intense enough to generate symptoms or other phenotypic manifestations of psychopathology.

An investigation by Reyher and Shoemaker (1961) is also pertinent. TAT cards were utilized as stimuli for producing age regressions and the reliving of important emotionalized experiences. Ten TAT cards were selected randomly to be conflictual or neutral for four subjects who were capable of deep hypnosis.

In order to create a conflict to each of five cards, hypnotized subjects were told, as they looked at each card, that disturbing emotions would be aroused. The subjects were then regressed to a time when these emotions were difficult to manage. In order to create nonconflictual or neutral reactions to the other five cards, the instructions were the same as above except that the emotions were nondisturbing. A posthypnotic amnesia was suggested and, in addition, the subject was told that the cards would stir up the same feelings as before, and that he would reveal them directly or indirectly in the stories that he would be asked to tell. In the waking state, the subject was given the same cards, by another experimenter, with standard instructions.

Although no symptoms were reported by the subjects marked differences were observed between the content of the hypnotic reactions and the waking stories. The conflict-cards were generally associated with more alterations than were the neutral-cards; however, on some occasions the latter also were associated with marked differences. These differences for both kinds of cards almost always reflected unresolved conflicts and helped guide psychotherapy.

Unfortunately the paradigm had not been formulated before this investigation was carried out. The induced processes were not kept distinct from suggested phenotypic be-

havior, as the instructions do not permit the induced process to generate spontaneously all of the subject's behavior: the subject is instructed to tell a story related directly or indirectly to the induced process. Since, in broad terms, the subject is told how to respond, it is impossible to determine what responses were generated by the induced process and what responses were a direct reflection of the hypnotic suggestion. In order to satisfy the paradigm, the instructions should state that the posthypnotic administration of each TAT card will stir up the same thoughts and feelings as it did during the hypnosis and that these reactions will become overwhelmingly intense.

Subsequent experience has shown that the indirect versus direct option can be dropped, because even neurotic subjects are not easily overwhelmed by hypnotically induced conflict. The subject's defensive organization usually does a good job in regulating hypnotically induced stress; nevertheless, the experimenter must be constantly alert for signs of a serious breakdown in ego functions when the subject is experiencing distress.

The absence of overt psychopathology may be attributed to the fact that the impulse was not suggested to be overwhelming and that the subject was given an option of how to respond. Despite these inadequacies in design for the production of psychopathology, the hypnotic reactions that produced the most alterations in the waking stories were congruent with central areas of conflict described in earlier psychodiagnostic impressions but which had not yet come up in psychotherapy. This observation indicates that there were significant psychodynamic reactions involved and that this procedure might be very productive of psychopathology, if utilized properly.

There is reason to believe that the conflict-producing potential of the hypnotic reactions can be intensified. It was observed that in subsequent psychotherapeutic sessions with the subjects, "deeper" aspects of the hypnotic reactions which were markedly changed in the waking state often could be uncovered by the induction of successive dreams about the material "behind" them. By telling the subject that he would have a dream about the emotions and thoughts behind his hypnotic experience, the material often became progressively more clearly represented until an abreaction of emotionally charged experiences took place. This material is valuable from the point of view of psychotherapy and, for research purposes, may be used to produce anxiety and psychopathology in the posthypnotic state.

These procedures would seem to have the greatest potential for creating psychopathology, but they also have the disadvantage that they should be restricted to subjects who are in psychotherapy or those who are waiting to begin. The experimenter must be in a position to help the subject work out adverse reactions if they should occur; otherwise, he places himself in an untenable ethical and professional position.

Discussion

All previous research in the hypnotic induction of psychopathology in some way has interfered with the spontaneous reactions of the subjects by instructing them how to react to the induced processes; consequently, the interpretative significance of the subjects' reactions is reduced in proportion to the extent of the interfer-

ence. If the induced processes have no intrinsic capacity for spontaneously producing alterations in behavior—such as distortions, repression, psychosomatic reactions, etc.—then the induced processes have no real clinical significance, and the imposed experimental reactions are merely hypnotic suggestions to be carried out.

Two methods derived from psychoanalytic theory were presented in which emotions can be linked with anxiety and the production of psychopathology: one artificial and the other natural. First, an emotion, such as hate, is brought to *overwhelming* intensity by a set of appropriate hallucinatory experiences (paramnesia). Since the intense hate would pose a vital threat to the subject's security under the circumstances of the waking state, it is hypothesized that its activation by a posthypnotic signal creates the conditions for conflict, anxiety, and psychopathology. The posthypnotic intensification of hostility activates the subject's traditional defenses against hostility of such intensity; that is, there is a danger point in the intensity of hostility beyond which the subject would lose control and, thereby, subject himself to the retaliation of the environment. The necessary controls and defenses are learned early in life and are triggered off in the posthypnotic state at the time the relevant posthypnotic signal is given. The second method is the same as the first except that the subject's own idiosyncratic conflicts are activated by the posthypnotic signal. His defenses against anxiety-producing processes are pressed beyond their usual limits, and anxiety and psychopathology are produced.

The paradigm can be utilized to test theories regarding specific kinds of psychopathology and almost any alteration in personality. For example, if a state of depression is desired, there are at least two clinical models from which to choose: the subject is made to believe that the objects and symbols for the gratification of his important emotional needs are no longer available, and in the waking state, he is given a paramnesia consistent with these events; if clinical, reactive depression is desired, hostility is induced toward loved ones in subjects who, on the basis of previous knowledge, turn this kind of hostility inwards. More directly, it may be possible to condition subjects who have a tendency in this direction to react to their own hostility by turning it inwards, and then to produce a paramnesia which involves a situation that normally would lead to intense hostility. The subject is given a posthypnotic signal for this hostility to become intense and conscious. Only those subjects are retained for study who do not achieve awareness of their hostility, despite the posthypnotic suggestion to do so.

In regard to the induction of a paramnesia or the implantation of an impulse that ordinarily is foreign to the subjects, there is reason to believe that a suggested amnesia for the hypnotic session may be necessary. If an amnesia is not suggested, it may be that enough fragments of the session will be recalled by the subject for him to realize that the experimenter had implanted something, and the growth of subsequent insight into the true nature of the experience would render the conflict innocuous. A suggested amnesia would prevent the subject from acquiring insight and preserve the conflict.

This also may be true for the activation of the subject's own conflicts. The fact that the experimenter succeeds in getting a hypnotized sub-

ject to become aware of conflictual material indicates that repression already had started to break down and that the subject may find it relatively easy to become aware of the material in the posthypnotic state. The hypnotic uncovering of conflictual material in patients undergoing psychotherapy supports this observation. When potent repressed material becomes represented in hypnosis, this indicates that the forces maintaining repression have been growing progressively weaker. This is illustrated by the fact that only after many months of intensive psychotherapy, including hypnoanalytic techniques, do the most significant repressions begin to lift. They begin to break down because the way has been prepared by the progressive development of a more secure relationship with the psychotherapist and the prior achievement of insight into less intense facets of basic conflicts. Most psychotherapists who are experienced with hypnoanalytic techniques realize that while hypnosis is not an immediate and direct route to the uncovering of repressed material, it is certainly more rapid and more direct than most other methods. Once something has been uncovered in hypnosis, subsequent insight in the waking state is usually attained readily. It may be that a posthypnotic amnesia reinforces repressive forces and thereby preserves the capacity of the induced dynamics to produce psychopathology.

There is some evidence that the relationship between the experimenter and the subject is a significant factor in successfully inducing conflicts. In an unpublished study, the author was able to replicate the results of an earlier study (Reyher, 1961) which produced somatic and psychological reactions to the posthypnotic stimulation of hypnotically induced conflict. However, an assistant using the same procedure could produce only symptoms of a relatively mild degree. Other than the different experimenters, there was one obvious difference in the preparation of the subjects that might account for this discrepancy. The subjects who were used by the assistant were hypnotized and brought to a deep trance by someone else. The assistant merely saw them for one brief session before the experimental session in order to establish the depth of the trance. In contrast, the senior experimenter had begun with naive subjects and brought them to a deep level of hypnosis himself in the course of three or four sessions. By the time the subjects were ready for the experimental sessions, they seemed to be quite at ease and trusting of the experimenter. It is reasonable to hypothesize that an unfamiliar experimenter would arouse some anxiety and defensive behavior that would interfere with the effect of any suggestions of a personal nature.

No matter what is to be induced hypnotically, it is wise to present the instructions in the passive voice. The use of the passive voice reduces the possibility that the subject may act out a role to please the experimenter. More specifically, the subject should not be instructed to carry out suggestions but he should be informed that he will be acted upon by something or that he is going to experience something. Not only does the active voice promote the expectation that the subject should do something, but it also enhances volitional, adaptive processes which render the hypnotic behavior similar to waking behavior. Thus, the instructions should minimize the role of volitional processes and maximize the role of nonvolitional processes.

REFERENCES

ADAMS, J. K. Laboratory studies of behavior without awareness. *Psychol. Bull.*, 1957, **54**, 383–405.

AINSWORTH, M. Problems of validation. In B. Klopfer, M. Ainsworth, W. Klopfer, and R. Holt (Eds.), *Developments in the Rorschach technique.* Vol. I. New York: Yonkers-on-Hudson, 1954. Pp. 405–500.

BOBBITT, R. A. The repression hypothesis studied in a situation of hypnotically induced conflict. *J. abnorm. soc. Psychol.*, 1958, **56**, 204–212.

COUNTS, R. M., & MENSH, I. N. Personality characteristics in hypnotically induced hostility. *J. clin. Psychol.*, 1950, **6**, 325–330.

EISENBUD, J. Psychology of headache. *Psychiat. Quart.*, 1937, **11**, 592–619.

ERICKSON, M. H. The method employed to formulate a complex story for the induction of an experimental neurosis in a hypnotic subject. *J. gen. Psychol.*, 1944, **31**, 67–84.

GORDON, J., BARCLAY, M., & LUNDY, M. Galvanic skin responses during repression, suppression, and verbalization in psychotherapeutic interviews. *J. consult. Psychol.*, 1959, **23**, 243–251.

HUSTON, P. E., SHAKOW, D., & ERICKSON, M. H. A study of hypnotically induced complexes by means of the Luria technique. *J. gen. Psychol.*, 1934, **11**, 65–97.

LEVITT, E. E., DEN BREEIJEN, A., & PERSKY, H. The induction of clinical anxiety by means of a standardized technique. *Amer. J. clin. Hypnosis*, 1960, **2**, 206–214.

LURIA, A. R. *The nature of human conflict.* New York: Liverright, 1932.

ORNE, M. T. The nature of hypnosis: Artifact and essence. *J. abnorm. soc. Psychol.*, 1959, **58**, 277–299.

REYHER, J. Posthypnotic stimulation of hypnotically induced conflict in relation to psychosomatic reactions and psychopathology. *Psychosom. Med.*, 1961, **23**, 384–391.

REYHER, J., & SHOEMAKER, D. A comparison between hypnotically induced age regressions and waking stories to TAT cards: A preliminary report. *J. consult. Psychol.*, 1961, **25**, 409–413.

ROSEN, H. *Hypnotherapy in clinical psychiatry.* New York: Julian Press, 1953.

WEITZENHOFFER, A. M. *Hypnotism: An objective study in suggestibility.* New York: Wiley, 1953.

WOLBERG, L. R. Hypnotic experiments in psychosomatic medicine. *Psychosom. Med.*, 1947, **9**, 337–342.

(Received May 10, 1961)

GEORGE BANTA COMPANY, INC., MENASHA, WISCONSIN, U.S.A.